Testing for Unit Roots in Panels by Using a Mixture Model

Madsen, Edith

Publication date:
2003

Document version
Publisher's PDF, also known as Version of record

CAM

Centre for Applied Microeconometrics

Institute of Economics
University of Copenhagen

http://www.econ.ku.dk/CAM/

Testing for unit roots in panels by using a mixture model

Edith Madsen

2003-10

The activities of CAM are financed by a grant from
The Danish National Research Foundation

Testing for unit roots in panels by using a mixture model

Edith Madsen[†]

Institute of Economics, University of Copenhagen, Denmark

August 2003

Abstract

This paper introduces a dynamic panel data model where the regression coefficients are allowed to vary across cross-section units. The framework is a mixture model obtained by mixing two dynamic panel data models with different parameters according to some mixing weights. The parameters in the model are estimated by the method of maximum likelihood, and it is shown that the maximum likelihood estimator is consistent and asymptotically normal. Within the mixture model it is possible to distinguish between different unit root hypotheses that cannot be distinguished by existing test procedures. More specifically, it is possible to test the hypothesis that a group of cross-section units has time-series processes with a unit root. The method is applied to income data from the PSID. For this sample there is no evidence of unit roots in the income processes but the processes differ substantially between individuals.

Keywords: Dynamic panel data model; Mixture model; Maximum likelihood estimation; Random coefficients; Unit roots

JEL classification: C13; C16; C23

[†] Institute of Economics, University of Copenhagen, Studiestræde 6, 1455 Kbh. K., Denmark
Email: edith.madsen@econ.ku.dk

1 Introduction

In this paper we consider estimation of the parameters in a dynamic panel data model where the regression coefficients differ across cross-section units. We consider traditional micro-panels where the time-series dimension is small relative to the cross-section dimension. The model used in this paper can be considered as being a compromise between two extreme versions of the dynamic panel data model. One extreme is when the regression coefficients are the same for all cross-section units and the other extreme is when the regression coefficients vary completely at random across cross-section units. In between are versions of the model where the variation in the regression coefficients is described by a specific distribution. In this paper the joint distribution of all parameters, including the regression coefficients, is assumed to be discrete with only two possible values of the parameters. The model can be interpreted as resulting from mixing two groups of cross-section units where each group is characterized by the value of the parameters describing their time-series processes. The mixing weights describing the probability that a given cross-section unit belongs to either one of the two groups are also allowed to vary across cross-section units. Again the variation is restricted such that cross-section units with the same observable characteristics have the same mixing weights.

The main contribution of the paper is to show that the maximum likelihood estimator of the parameters in the mixture model described above has the usual asymptotic properties, i.e. it is consistent and asymptotically normal. This result differs from the one usually found in this type of mixture model, since in general the maximum likelihood estimator of the parameters does not exist, see Titterington, Smith & Markov (1985) and McLachlan & Peel (2000). It turns out that when the model contains exogenous variables it is necessary to impose a restriction between the number of exogenous variables and the time-series dimension of the panel in order to show the result. In particular, this restriction rules out the possibility of including time dummies for each period of time in the model.

It is well-known that in dynamic micro-panels where the regression coefficients differ across cross-section units, inference on the mean coefficients based on standard panel data methods can be very misleading, see Robertson & Symons (1992) and Pesaran & Smith (1995). In spite of this, only few alternative inference procedures have been suggested. Alvarez, Browning & Ejrnæs (2002) investigate the topic in the modelling of income processes for individuals. As the title of their paper states, they allow for 'lots of heterogeneity' meaning that most of the parameters appearing in their dynamic panel data model vary across individuals according to some specific distributions. Estimation of the parameters is done by using a simulated minimum distance procedure. There are two main differences between their approach and the approach used in this paper. First, their model does not allow for the inclusion of any exogenous variables. Second, the parameters in their model, such as the regression coefficients and the error variances, are assumed to be distributed independently of each other. In the model used in this paper, the parameters are allowed to be correlated and no parametric form is imposed on the dependency between the regression coefficients and the error variances. Hence, our approach is more in line with semi-parametric modelling, see Section 1.4 in McLachlan & Peel (2000). In addition, Alvarez, Browning

& Ejrnæs (2002) emphasize that their choice of parameter distributions is only appropriate for their specific data set. Conversely, the method suggested in this paper can be applied to any data set and hence it can be considered as a general tool. Hsiao, Pesaran & Tahmscioglu (1999) consider estimation of the mean regression coefficients in a dynamic panel data model by using a Bayesian approach. However, in relation to unit root inference the mean regression coefficient is not that interesting.

Mixture models were first introduced into econometrics by Quandt (1972) and within this field they are often referred to as switching regression models. This type of model has not been used in the analysis of non-linear dynamic panels whereas it is well-known in the analysis of non-linear time series. For example, Wong & Li (2000), Wong & Li (2001) and Rahbek & Shephard (2002) consider mixtures of autoregressive time-series models. These models are used to describe variables with the property that their behavior changes over time. Even though the interpretation of these models is very different from the model considered in this paper, the paper by Rahbek & Shephard (2002) has been a source of inspiration.

The mixture model considered in this paper makes it possible to distinguish between unit root hypotheses that can not be distinguished by any of the existing test procedures. Let us shortly review the existing test procedures. Breitung & Meyer (1994), Breitung (1997) and Harris & Tzavalis (1999) all consider testing the null hypothesis of all cross-section units having autoregressive time-series processes with an autoregressive coefficient of unity. The alternative hypothesis is that all cross-section units have stationary autoregressive time-series processes with the same autoregressive coefficient. The test statistics are all based on Least Squares (LS) or Generalized Method of Moments (GMM) estimators of the autoregressive coefficient. Im, Pesaran & Shin (2003) consider testing the same null hypothesis but against an alternative hypothesis where the autoregressive coefficient differs across cross-section units and is less than unity for at least a group of the cross-section units. Hence, the alternative hypothesis includes the case where a group of the cross-section units have unit root processes while the others have stationary processes with the same autoregressive coefficient. Their test statistic is based on the cross-section average of individual Dickey-Fuller test statistics which are obtained separately for each cross-section unit. A common feature of these tests is that they test the null hypothesis of all cross-section units having an autoregressive coefficient of unity. But if this hypothesis is rejected the tests do not provide any information on why that is. The finding could be the result of the fact that only a group of the cross-sections units has an autoregressive coefficient of unity. In that case, it is interesting to know the size of this group and also to know which cross-section units are most likely to belong to this group. Or it could be the result of the fact that none of the cross-section units have an autoregressive coefficient of unity. Within the mixture model considered in this paper, it is possible to distinguish between these explanations as they can be formulated as hypotheses on the parameters in the model.

The rest of the paper is organized as follows. In Section 2 the mixture model and the underlying assumptions are specified. In Section 3 the estimation procedure is discussed. In Section 4 unit root testing is discussed. In Section 5 the method is applied to income data from the PSID. Section 6 provides

some concluding remarks.

The following notation is used throughout the paper. The symbol $\xrightarrow{P}$ denotes convergence in probability and $\xrightarrow{w}$ denotes weak convergence. The matrix norm $\|A\|$ is defined as $\|A\| = \mathrm{tr}\{A'A\}^{\frac{1}{2}}$. The inequality $A > 0$ means that the matrix A is positive definite.

2 The mixture model

We consider the two-component mixture model defined by

$$\begin{array}{lll} y_{it} = \rho_1 y_{it-1} + z'_{it}\omega_1 + \alpha_{1i} + \varepsilon_{1,it} & \text{for } t = 1,...,T & \text{with probability } p_i \\ y_{it} = \rho_2 y_{it-1} + z'_{it}\omega_2 + \alpha_{2i} + \varepsilon_{2,it} & \text{for } t = 1,...,T & \text{with probability } 1-p_i \end{array} \quad \text{for } i = 1,...,N \qquad (1)$$

where $\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are iid normal across i and t with means zero and variances $\sigma^2_{1\varepsilon}$ and $\sigma^2_{2\varepsilon}$, and the terms α_{1i} and α_{2i} represent individual-specific effects that are unobserved. For notational convenience we assume that the initial values y_{i0} are observed. First of all, let us consider the situation where there are no exogenous variables in the model, i.e. $z_{it} = 0$. Without loss of generality we assume that $\rho_1 \leq \rho_2$ and we refer to the process defined by the first (second) equation as a low-persistency (high-persistency) process. Using this terminology, the model expresses that given the initial value y_{i0} then y_{it} is generated by a low-persistency process with probability p_i and by a high-persistency process with probability $1-p_i$. A property of this mixture model is that the mixing is done solely in the cross-section dimension not in the time-series dimension. It means that for a given cross-section unit the parameters are constant over time. This property appears to be important in relation to the likelihood analysis discussed in Section 3. The model can also be interpreted as resulting from mixing two groups of individuals where each group is characterized by the values of the AR coefficient ρ, the error variance σ^2_ε and the individual-specific term α_i.

Returning to the model with exogenous variables, z_{it} is a $k \times 1$ vector of such variables which can be time-constant as well as time-varying including a constant term. The variable z_{it} can also contain lags of the exogenous variables but this is left implicit as it does not affect the estimation procedure as long as z_{it} is strictly exogenous. However, with respect to interpretation of the model in relation to unit root hypotheses the specific form of z_{it} matters. This is discussed in detail in Section 4. The two $k \times 1$ vectors of parameters ω_1 and ω_2 reflect that the variable z_{it} can affect individuals in the two groups differently. For instance, the two groups of individuals can have different levels and different linear trends. So at first sight, the model seems to allow for a lot of heterogeneity between cross-section units. However, in relation to the likelihood analysis in Section 3 it appears that there is restriction between the number of exogenous variables and the time-series dimension of the panel. In particular, it is not possible to include time-dummies for each period of time in the model no matter what the time-series dimension is. This is a common problem in this type of non-linear dynamic panel data model. For example, the method suggested by Im, Pesaran & Shin (2003) to test for unit roots in the time-series processes for univariate variables does not allow for time dummies and the method suggested by Alvarez, Browning & Ejrnæs (2002) to model heterogeneity in income dynamics does not allow time-dummies or any other

exogenous variables to be included. Thus, an important feature of the model defined by (1) is that it allows for the inclusion of *some* exogenous variables and heterogeneity through these.

The mixing weights describing the mixing proportions of the two components in (1) are allowed to vary across units. We assume that this variation is described by a logistic function of some time-constant variables. This is expressed in the following way

$$p_i = \frac{\exp(\gamma' D_i)}{1 + \exp(\gamma' D_i)} \tag{2}$$

where D_i is a $m \times 1$ vector of random variables that are constant over time with its first element equal to 1, and γ is a $m \times 1$ vector of parameters. When $\gamma = (\gamma_1, 0, ..., 0)'$ the mixing weights are the same for all cross-section units, i.e. $p_i = p$ for $i = 1, ..., N$.

To specify the model defined by (1) and (2) further, we impose the following assumptions.

Assumption 1 (***General assumptions***)

(i) *$\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are both iid normal across i, t with means zero variances $\sigma^2_{1\varepsilon}$ and $\sigma^2_{2\varepsilon}$*

(ii) *$\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are both independent of α_{1i} and α_{2i} for all $t = 1, ..., T$*

(iii) *$\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are both independent of y_{i0}, D_i and z_{is} for all $t, s = 1, ..., T$*

Assumption 1 states that the errors $\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are iid normal across i, t and independent of all other terms in the process generating y_{it}. In particular, $\varepsilon_{1,it}$ and $\varepsilon_{2,it}$ are independent of z_{is} for all $s = 1, ..., T$, i.e. z_{it} is strictly exogenous. Note that the assumption does not impose any restrictions on the relation between the individual-specific terms $(\alpha_{1i}, \alpha_{2i})$ and y_{i0}, D_i and z_{it} for $t = 1, ..., T$.

Assumption 2 (***Additional assumptions***)

(i) y_{i0} is iid across i with finite sixth order moments, i.e. $E|y_{i0}|^6 < \infty$

(ii) D_i is iid across i with finite sixth order moments, i.e. $E\|D_i\|^6 < \infty$

(iii) z_{it} is iid across i with finite sixth order moments, i.e. $E\|z_{it}\|^6 < \infty$ for all $t = 1, ..., T$

Assumption 2 states that the variables y_{i0}, D_i and z_{it} for $t = 1, ..., T$ are all iid across i with finite sixth order moments. Here the assumption about the variables being distributed independently across i is crucial in relation to the likelihood analysis while the assumption about the variables being distributed identically across i is not. The latter assumption is imposed in order to simplify the likelihood analysis but can easily be dropped. In that case, slightly stronger moment conditions are required, see Section 4.2 in Amemiya (1985).

Before discussing how to treat the unobserved individual-specific effects α_{1i} and α_{2i} we introduce some notation. We let y_i and $y_{i,-1}$ be the $T \times 1$ vectors defined as $y_i = (y_{i1}, ..., y_{iT})'$ and $y_{i,-1} = (y_{i0}, ..., y_{iT-1})'$, z_i be the $T \times k$ matrix defined as $z_i = \begin{bmatrix} z_{i1} & z_{i2} & \cdots & z_{iT} \end{bmatrix}'$ and ι_T be a $T \times 1$ vector of ones. In addition we let ζ denote the vector of parameters in the model defined by (1) and (2) conditional on α_{1i} and α_{2i}, i.e. $\zeta = \left(\rho_1, \omega_1', \sigma^2_{1\varepsilon}, \rho_2, \omega_2', \sigma^2_{2\varepsilon}, \gamma'\right)'$. Then for every $i = 1, ..., N$ the density function of y_i

conditional on the variables y_{i0}, D_i and z_i and the individual-specific terms α_{1i} and α_{2i} is the following except for a constant

$$\tilde{f}_\zeta\left(y_i\middle|\, y_{i0}, D_i, z_i, \alpha_{1i}, \alpha_{2i}\right) \propto p_i \tilde{\phi}_i\left(\rho_1, \omega_1, \sigma_{1\varepsilon}^2\middle|\, \alpha_{1i}\right) + (1-p_i)\,\tilde{\phi}_i\left(\rho_2, \omega_2, \sigma_{2\varepsilon}^2\middle|\, \alpha_{2i}\right) \tag{3}$$

where

$$\tilde{\phi}_i\left(\rho, \omega, \sigma_\varepsilon^2\middle|\, \alpha_i\right) = \left(\sigma_\varepsilon^2\right)^{-\frac{T}{2}} \exp\left\{-\left(2\sigma_\varepsilon^2\right)^{-1}\left(y_i - \rho y_{i,-1} - z_i\omega - \alpha_i \iota_T\right)'\left(y_i - \rho y_{i,-1} - z_i\omega - \alpha_i \iota_T\right)\right\} \tag{4}$$

Now it is not possible to estimate the parameter ζ by using the density above as the terms α_{1i} and α_{2i} are unobserved. There are two common ways to deal with unobserved individual-specific effects. The first approach treats α_{1i} and α_{2i} as nuisance parameters. In this case, the number of parameters increases with the number of observations in the cross-section dimension N. This is the classic incidental-parameter problem first discussed in the paper by Neyman & Scott (1948). The second approach avoids this problem by treating $(\alpha_{1i}, \alpha_{2i})$ as being a random variable with a common distribution function G (conditional on y_{i0}, D_i and z_i) for all $i = 1, ..., N$. In this case the density function of y_i conditional on y_{i0}, D_i and z_i is the following

$$f_{\zeta,G}\left(y_i\middle|\, y_{i0}, D_i, z_i\right) = \int \tilde{f}_\zeta\left(y_i\middle|\, y_{i0}, D_i, z_i, \alpha_{1i}, \alpha_{2i}\right) dG\left(\alpha_{1i}, \alpha_{2i}\middle|\, y_{i0}, D_i, z_i\right) \tag{5}$$

If no parametric assumptions are imposed on the distribution function G the density above specifies a semiparametric mixture model for y_i. Kiefer & Wolfowitz (1956) give conditions that ensure consistent estimation of the parameters ζ and the distribution function G. However, in this paper we impose a parametric form on G which allows for correlation between the individual-specific effects and the initial value. This is done by imposing the assumption below.

Assumption 3 (***Unobserved individual-specific effects***)
Conditional on y_{i0}, D_i and z_i the terms α_{1i} and α_{2i} are independent of each other with the following conditional distributions

$$\begin{aligned}
\left(\alpha_{1i}\middle|\, y_{i0}, D_i, z_i\right) &\sim N\left(\alpha_1 y_{i0}, \sigma_{1\alpha}^2\right) \quad \textit{where} \;\; \sigma_{1\alpha}^2 \geq 0 \\
\left(\alpha_{2i}\middle|\, y_{i0}, D_i, z_i\right) &\sim N\left(\alpha_2 y_{i0}, \sigma_{2\alpha}^2\right) \quad \textit{where} \;\; \sigma_{2\alpha}^2 \geq 0
\end{aligned}$$

This specification of the individual-specific effects is suggested by Chamberlain (1980) in a dynamic panel data model without heterogeneity in the AR coefficient, see also Blundell & Smith (1991), Blundell & Bond (1998) and Wooldridge (2002a). Compared to the assumptions underlying a standard random effects model, Assumption 3 is much less restrictive as it allows for correlation between the individual-specific effects and the initial values. To get a better understanding of the assumption, we consider the situation where the variable y_{it} is income of individual i at time t and we consider two individuals with the same initial values. If the two individuals are from different groups, the assumption can be interpreted as allowing the expected values of the individual-specific effects, reflecting such things as ability and motivation of individuals, to be different. On the other hand, if the two individuals are from the same group the expected values of the individual-specific effects are the same.

It is also possible to let the conditional mean of the individual-specific effects depend on the exogenous variable z_i, for example by letting $E(\alpha_{1i}|\, y_{i0}, D_i, z_i) = \alpha_j y_{i0} + \bar{z}_i'\pi_j$ for $j = 1, 2$ where $\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$ and π_1, π_2 are $k \times 1$ vectors of parameters. If z_{it} is constant over time this specification does not change the model but it means that only the sum of the parameters ω_j and π_j for $j = 1, 2$ is identified. In the empirical example with income considered in Section 5, the exogenous variables are functions of age and education level of the individual. In this case, it is likely that the education level (which is constant over time in our sample) is affected by the individual-specific effects while the age is not. Taking this into account is left implicit since it does not change the model.

Defining $\varepsilon_{j,i0} = \alpha_{ji} - \alpha_j y_{i0}$ for $j = 1, 2$, the model in (1) can be rewritten as follows

$$\begin{aligned} y_{it} &= \rho_1 y_{it-1} + z_{it}'\omega_1 + \alpha_1 y_{i0} + \varepsilon_{1,i0} + \varepsilon_{1,it} \quad \text{for } t = 1, ..., T \quad \text{with probability } p_i \\ y_{it} &= \rho_2 y_{it-1} + z_{it}'\omega_2 + \alpha_2 y_{i0} + \varepsilon_{2,i0} + \varepsilon_{2,it} \quad \text{for } t = 1, ..., T \quad \text{with probability } 1 - p_i \end{aligned} \tag{6}$$

where $(\varepsilon_{j,it}|\, y_{i0}, D_i, z_i)$ is iid $N\left(0, \sigma_{j\varepsilon}^2\right)$ across i, t and $(\varepsilon_{j,i0}|\, y_{i0}, D_i, z_i)$ is iid $N\left(0, \sigma_{j\alpha}^2\right)$ across i for $j = 1, 2$. In addition, $(\varepsilon_{j,it}|\, y_{i0}, D_i, z_i)$ and $(\varepsilon_{j,i0}|\, y_{i0}, D_i, z_i)$ are independent of each other for all $t = 1, ..., T$ according to Assumption 1. This means that $(v_{j,it}|\, y_{i0}, D_i, z_i)$ where $v_{j,it} = \varepsilon_{j,i0} + \varepsilon_{j,it}$ for $j = 1, 2$ satisfies the assumption being made about the error term in a standard random effects model. We define the $(k+2) \times 1$ vectors ξ_1 and ξ_2 as $\xi_j = \left(\rho_j, \omega_j', \alpha_j\right)'$ for $j = 1, 2$, and the $T \times (k+2)$ matrix Z_i as $Z_i = \left[\begin{array}{ccc} y_{i,-1} & z_i & y_{i0}\iota_T \end{array}\right]$. Then with $\vartheta = \left(\xi_1', \sigma_{1\varepsilon}^2, \sigma_{1\alpha}^2, \xi_2', \sigma_2^2, \sigma_{2\alpha}^2, \gamma'\right)'$ the density function of y_i conditional on y_{i0}, D_i and z_i is the following except for a constant

$$f_\vartheta\left(y_i|\, y_{i0}, D_i, z_i\right) \propto p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, \sigma_{1\alpha}^2\right) + (1 - p_i)\,\phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, \sigma_{2\alpha}^2\right) \tag{7}$$

where

$$\phi_i\left(\xi, \sigma_\varepsilon^2, \sigma_\alpha^2\right) = |V|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(y_i - Z_i\xi\right)' V^{-1}\left(y_i - Z_i\xi\right)\right\} \tag{8}$$

with $V = \sigma_\varepsilon^2 I_T + \sigma_\alpha^2 \iota_T \iota_T'$ where I_T is the $T \times T$ identity matrix and $\iota_T\iota_T'$ a $T \times T$ matrix of ones. This means that when including the initial value y_{i0} as an additional regressor in each of the T equations, the density of each component in the mixture model defined by (8) corresponds to that of a standard random effects model.

The likelihood analysis in Section 3 is done conditional on y_{i0}, D_i and z_i. Here the choice of conditioning on the initial values y_{i0} is probably the most controversial. Therefore, we end this section with a short discussion of this issue, see also Section 13 in Wooldridge (2002b).

The distribution of y_{i0} (conditional on D_i and z_i) will typically depend on the parameter ϑ. In spite of this, the conditional maximum likelihood estimator of ϑ is consistent but it might not be efficient. Clearly, this statement is only true if the conditional model is correctly specified. On the other hand, consistency of the unconditional maximum likelihood estimator of the parameters also requires a correct specification of the distribution of y_{i0}. No matter which approach we choose, it involves specification of the relation between y_{i0} and the individual-specific effects. When the AR coefficient is less than unity in absolute value it is obvious to specify the initial values such that the time-series processes become covariance stationary, see e.g. Bhargava & Sargan (1983). Clearly, this approach can not be used in

this paper since we are testing the hypothesis that an AR coefficient equals unity. Hsiao, Pesaran & Tahmiscioglu (2002) suggest another approach where the model is expressed in first-diffences in order to eliminate the individual-specific effects. To obtain an expression for the joint density of $\Delta y_{i1}, ..., \Delta y_{iT}$ (only conditional on possible exogenous variables) they impose some restrictions on the initial changes Δy_{i1} and on the behavior of possible exogenous variables. Most importantly, these restrictions do not require that the AR coefficient is less than unity in absolute value. So in principle, their idea could be incorporated within the framework considered in this paper. In that case, the joint density of $\Delta y_{i1}, ..., \Delta y_{iT}$ will be on the same form as in (7) but with a different component density. In future work it would be interesting to compare the approach suggested by Hsiao, Pesaran & Tahmiscioglu (2002) to the one used in this paper.

3 Maximum likelihood inference

In this section we consider maximum likelihood estimation of the parameters in the mixture model specified in the previous section. We derive the asymptotic properties of this estimator under the assumption that $N \to \infty$ and T is fixed.

3.1 Estimation

First of all, the covariance matrix V appearing in the component density ϕ_i in equation (8) is parametrized in terms of σ_ε^2 and q where $q = \sigma_\alpha^2/\sigma_\varepsilon^2$. Clearly, there is a one-to-one correspondence between $\left(\sigma_\varepsilon^2, q\right)$ and $\left(\sigma_\varepsilon^2, \sigma_\alpha^2\right)$. Defining $J_T = \iota_T \iota_T'/T$ and $C_T = I_T - J_T$ which are the between-group transformation matrix and the within-group transformation matrix, respectively, the results below are obtained, e.g. Section 2.3 in Baltagi (1995).

$$V = \sigma_\varepsilon^2 I_T + \sigma_\alpha^2 \iota_T \iota_T' = \left(T\sigma_\alpha^2 + \sigma_\varepsilon^2\right) J_T + \sigma_\varepsilon^2 C_T \tag{9}$$

$$V^{-1} = \left(T\sigma_\alpha^2 + \sigma_\varepsilon^2\right)^{-1} J_T + \left(\sigma_\varepsilon^2\right)^{-1} C_T = \left(\sigma_\varepsilon^2\right)^{-1} \left(\left(1 + Tq\right)^{-1} J_T + C_T\right) \tag{10}$$

$$|V|^{-1} = \left(\sigma_\varepsilon^2\right)^{-(T-1)} \left(T\sigma_\alpha^2 + \sigma_\varepsilon^2\right)^{-1} = \left(\sigma_\varepsilon^2\right)^{-T} \left(1 + Tq\right)^{-1} \tag{11}$$

We note that this parametrization can also be used when the panel data set is unbalanced such that each cross-section unit i is observed for T_i time periods. This is not the case when V is parametrized in terms of σ_ε^2 and $\sigma_\varepsilon^2/(T\sigma_\alpha^2 + \sigma_\varepsilon^2)$ as often seen in the literature on maximum-likelihood estimation of random effects models, e.g. Section 2.4 in Baltagi (1995). In this case, the parametrization of V will be different for cross-section units with different T_i and therefore it can not be used. Thus, in order to make the results provided in this section applicable to unbalanced panel data sets such as the data set considered in Section 5, we parametrize V in terms of σ_ε^2 and q.

The freely varying parameters of the model defined by the equations (1) and (2) are given by $\vartheta = (\xi_1', \sigma_{1\varepsilon}^2, q_1, \xi_2', \sigma_{2\varepsilon}^2, q_2, \gamma')'$ where ξ_1, ξ_2 are $(k+2) \times 1$ vectors, $\sigma_{1\varepsilon}^2, \sigma_{2\varepsilon}^2 > 0$, $q_1, q_2 \geq 0$ and γ is a $m \times 1$ vector. According to (7), (8) and the results above, the density function of y_i conditional on y_{i0}, D_i and

z_i is the following except for a constant

$$f_\vartheta\left(y_i|\, y_{i0}, D_i, z_i\right) \propto p_i\phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1-p_i\right)\phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right) \tag{12}$$

where

$$\phi_i\left(\xi, \sigma_\varepsilon^2, q\right) = \left(\sigma_\varepsilon^2\right)^{-\frac{T}{2}}\left(1+Tq\right)^{-\frac{1}{2}}\exp\left\{-\left(2\sigma_\varepsilon^2\right)^{-1}\left(y_i - Z_i\xi\right)'\left(\left(1+Tq\right)^{-1}J_T + C_T\right)\left(y_i - Z_i\xi\right)\right\} \tag{13}$$

For every $i = 1, ..., N$ the log-likelihood function conditional on y_{i0}, D_i and z_i is given by

$$l_i\left(\vartheta\right) = \log\left\{p_i\phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1-p_i\right)\phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)\right\} \tag{14}$$

Then by using the independency between cross-section units the conditional log-likelihood function can be written as

$$l_N\left(\vartheta\right) = \sum_{i=1}^N l_i\left(\vartheta\right) = \sum_{i=1}^N \log\left\{p_i\phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1-p_i\right)\phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)\right\} \tag{15}$$

It is well-known that in general the likelihood function for a mixture of normal distributions is unbounded, and so the maximum likelihood estimator of the parameters does not exist, see e.g. Redner & Walker (1984), Section 4.3 in Titterington, Smith & Markov (1985) and Section 3.8 in McLachlan & Peel (2000). This is not necessarily the case in the mixture model considered in the present paper. More specifically, it appears that the singularity problem does not occur when the number of exogenous variables is restricted in relation to the time-series dimension of the panel. To get a better understanding of why this is, we start out by describing the singularity problem within the most simple framework. This is done in Example 1 below.

Example 1 A mixture of two univariate normal distributions

Consider the situation where the univariate variable y_i is distributed according to the following mixture model

$$\begin{array}{ll} y_i = \mu_1 + \varepsilon_{1i} & \text{with probability } p \\ y_i = \mu_2 + \varepsilon_{2i} & \text{with probability } 1-p \end{array} \qquad \text{for } i = 1, ..., N \tag{16}$$

where ε_{1i} and ε_{2i} are independently distributed across i as $N\left(0, \sigma_1^2\right)$ and $N\left(0, \sigma_2^2\right)$, respectively, and where $\left(\mu_1, \sigma_1^2\right) \neq \left(\mu_2, \sigma_2^2\right)$ and $0 < p < 1$. The log-likelihood function for this model is given by

$$l_N\left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p\right) = \sum_{i=1}^N \log\left\{p\phi_i\left(\mu_1, \sigma_1^2\right) + \left(1-p\right)\phi_i\left(\mu_2, \sigma_2^2\right)\right\} \tag{17}$$

where

$$\phi_i\left(\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right\} \tag{18}$$

Now for any $i = 1, ..., N$ and any observation y_i it is possible to choose the parameter μ_1 such that $\left(y_i - \mu_1\right) = 0$. Then as $\sigma_1^2 \to 0$ the component density $\phi_i\left(\mu_1, \sigma_1^2\right) \to \infty$. For any other $j = 1, ..., N$ where $\left(y_j - \mu_1\right) \neq 0$ the component density $\phi_j\left(\mu_1, \sigma_1^2\right) \to 0$ as $\sigma_1^2 \to 0$, whereas the other component density $\phi_j\left(\mu_2, \sigma_2^2\right)$ is bounded away from zero. This means that there are N values of μ_1 where the

log-likelihood function tends to infinity on the boundary of the parameter space as the variance σ_1^2 tends to zero. In other words, one component of the mixture model can be used to obtain a perfect fit of one observation. In spite of this, the existence of a maximum likelihood estimator with the usual asymptotic properties (i.e. consistency and asymptotic normality) can be shown if the parameter space is restricted appropriately. Kiefer (1978) shows that this is the case when the parameter space is restricted to a compact neighborhood of the true parameter values. The corresponding estimator is referred to as a local maximum likelihood estimator. If the component variances are proportional, i.e. $\sigma_1^2 = k\sigma_2^2$ where $k > 0$ is known, the singularity problem does not occur. In relation to this, Hathaway (1985) shows that when the component variances are restricted such that $\sigma_1^2 \geq k\sigma_2^2$ for some $k > 0$, then the corresponding constrained maximum likelihood estimator has the usual asymptotic properties. Finally, Policello (1981) shows that the singularity problem does not occur when there are at least two different observations from each component in the mixture model.

First of all, we note that if parameter variation over time is also allowed it is not possible to avoid the singularity problem, as for any i, t it is always possible to choose the parameters ρ_1, ω_1 and α_1 such that $(y_{it} - \rho_1 y_{it-1} - z_{it}'\omega_1 - \alpha_1 y_{i0}) = 0$. Hence the assumption about the parameters being constant over time for a given unit is important. In our mixture model, the question is if for any value of (y_i, Z_i) it is possible to choose the parameter ξ_1 such that $(y_i - Z_i\xi_1) = 0$. From linear algebra we have the following result. For any value of y_i, the system $Z_i\xi_1 = y_i$ has at least one solution if and only if Z_i has full row rank. Negation of this statement yields the restriction on Z_i required in order to avoid the singularity problem. The restriction is that the number of rows in Z_i which equals T must be greater than the number of linearly independent columns in Z_i, i.e. T must be greater than the rank of Z_i. This means that time-dummies for each period of time can not be included in the model. Instead time-effects must be modelled such that the restriction on the number of exogenous variables is satisfied, for example by including linear time trends. The issue of which variables to include is discussed further in Section 4. In the simple model without exogenous variables and individual-specific effects, i.e. $\omega_j = 0$, $\alpha_{ji} = 0$ and $q_j = 0$ for $j = 1, 2$, the restriction is that $T \geq 2$. As each cross-section unit belongs to one of the two components, it means that there are at least two observations from each component of the mixture model. In this case, our result is related to the finding in Policello (1981), see Example 1 above.

Before we continue, the restriction is illustrated in Example 2 below.

Example 2 Consider the situation where $z_{it} = (1, d_i)'$ where d_i is a time-constant variable, for example a dummy variable. The $T \times 1$ vector y_i and the $T \times 4$ matrix Z_i are on the following forms

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} \qquad Z_i = \begin{bmatrix} y_{i0} & 1 & d_i & y_{i0} \\ \vdots & \vdots & \vdots & \vdots \\ y_{iT-1} & 1 & d_i & y_{i0} \end{bmatrix}$$

The rank of Z_i is 2 and hence if $T \geq 3$ it is not possible to find ξ such that $(y_i - Z_i\xi) = 0$ for any value of (y_i, Z_i).

Once the restriction described above is imposed, the maximum likelihood estimator of ϑ has the usual asymptotic properties. More specifically, it can be shown that there exists a sequence of roots of the likelihood equations which is consistent and asymptotically normally distributed. With probability tending to one as the sample size N tends to infinity, these roots correspond to a local maximum of the log-likelihood function. This is summarized in Theorem 1 below.

Theorem 1 *Consider the mixture model defined by the equations* (1) *and* (2)*. Under Assumption 1, 2, 3, when T is greater than the rank of Z_i, and when $\left(\xi_1', \sigma_{1\varepsilon}^2, q_1\right)' \neq \left(\xi_2', \sigma_{2\varepsilon}^2, q_2\right)'$, then there exists with probability tending to one as $N \to \infty$ a sequence $\hat{\vartheta}$ which satisfies the likelihood equations. The sequence $\hat{\vartheta}$ is a consistent estimator of the true parameter value ϑ, i.e.*

$$\hat{\vartheta} \xrightarrow{P} \vartheta \quad \text{as } N \to \infty \tag{19}$$

The limiting distribution of $\hat{\vartheta}$ is given by

$$\sqrt{N}\left(\hat{\vartheta} - \vartheta\right) \xrightarrow{w} N\left(0, I\left(\vartheta\right)^{-1}\right) \quad \text{as } N \to \infty \tag{20}$$

where

$$I\left(\vartheta\right) = E\left(\frac{\partial l_i\left(\vartheta\right)}{\partial \vartheta}\frac{\partial l_i\left(\vartheta\right)}{\partial \vartheta'}\right) = E\left(-\frac{\partial^2 l_i\left(\vartheta\right)}{\partial \vartheta \partial \vartheta'}\right) > 0 \tag{21}$$

The proof of Theorem 1 is given in Appendix A.3. It is based on verifying that the standard regularity conditions due to Cramér (1946) are satisfied. The assumption that $\left(\xi_1', \sigma_{1\varepsilon}^2, q_1\right)' \neq \left(\xi_2', \sigma_{2\varepsilon}^2, q_2\right)'$ means that at least one element in the vector $\left(\xi_1', \sigma_{1\varepsilon}^2, q_1\right)' - \left(\xi_2', \sigma_{2\varepsilon}^2, q_2\right)'$ is different from zero such that the two components in the mixture model are different. If the parameters in the two components are the same, the parameter γ describing the mixing weights is not identified. In this case, the information matrix $I\left(\vartheta\right)$ is singular such that the usual Taylor expansions of the maximum likelihood estimator are invalid. In particular, this means that likelihood-based test statistics, such as LR or LM statistics, of the hypothesis about the two components being the same do not have the usual asymptotic χ^2-distributions. In the literature this problem is expressed as a nuisance parameter, in this case γ, not being present under the null hypothesis, see Davies (1977), Davies (1987), Andrews & Ploberger (1994) and Hansen (1996). The issue of how to determine the number of components in a mixture model is not investigated in this paper. For a discussion of this problem see Section 6.3 in McLachlan & Peel (2000).

If the restriction on T in relation to the number of exogenous variables does not hold, for example because there are strong reasons for including time dummies for each time period in the model, the local maximum likelihood estimator of the parameters has the usual asymptotic properties, see Example 1. In practice, this usually means that the maximum likelihood estimate is obtained without imposing any restrictions on the parameter space. However, when the number of observations belonging to one component is low, it can be very difficult to determine if an estimate is spurious, see Section 3.10 in McLachlan & Peel (2000).

3.2 Optimization of the log-likelihood function

In principle, we could obtain the maximum likelihood estimator directly by numerically maximizing the log-likelihood function in (15) with respect to ϑ. However in practice, the maximization problem is simplified considerably by formulating it as an incomplete-data problem and applying the EM algorithm introduced by Dempster, Laird & Rubin (1977). The advantage of this approach is that it solely involves well-known optimization problems once the component density is well-known. In our case the component density defined in (13) corresponds to that of a standard random effects model.

Within an incomplete-data framework each observation y_i is interpreted as coming from one of the two components in the mixture model defined by (6) and the associated component-indicator s_i is unobserved. As our mixture model consists of two components s_i is binary where $s_i = 1$ if y_i comes from the first component of the mixture and $s_i = 0$ if y_i comes from the second component of the mixture. Furthermore, $\Pr(s_i = 1 | y_{i0}, D_i, z_i) = \Pr(s_i = 1 | D_i) = p_i$ where p_i is defined in (2). The joint density function of the complete data $y_i^C = (y_i, s_i)$ conditional on y_{i0}, D_i and z_i is the following except for a constant

$$\tilde{f}_\vartheta\left(y_i^C \middle| y_{i0}, D_i, z_i\right) \propto p_i^{s_i}\left(1-p_i\right)^{1-s_i} \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right)^{s_i} \phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)^{1-s_i}$$

Here the assumption about $\left(y_i^C \middle| y_{i0}, D_i, z_i\right)$ being independent across i is appropriate since it means that the distribution of the complete data implies the appropriate distribution of the incomplete (observed) data. Using this the complete-data log-likelihood function $l_C(\vartheta)$ is given by the following expression

$$l_C(\vartheta) = \sum_{i=1}^{N} s_i\left(\log p_i + \log \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right)\right) + \sum_{i=1}^{N}\left(1-s_i\right)\left(\log\left(1-p_i\right) + \log \phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)\right) \tag{22}$$

The EM algorithm then repeats the E-step and the M-step described below until convergence is achieved.

E-Step: This step consists of computation of the conditional expectation of the complete-data log-likelihood function $l_C(\vartheta)$ given the incomplete (observed) data $y = (y_1, ..., y_N)$ by using the current estimate of ϑ. As the complete-data log-likelihood function $l_C(\vartheta)$ is linear in the missing data $(s_1, ..., s_N)$ and $\left(y_i^C \middle| y_{i0}, D_i, z_i\right)$ is independent across i this corresponds to computation of the conditional expectation of s_i given the observed data for unit i, i.e.

$$p_i^* = E\left[s_i | y_{i0}, D_i, z_i, y_i\right] = \Pr\left(s_i = 1 | y_{i0}, D_i, z_i, y_i\right) = \frac{p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right)}{p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1-p_i\right)\phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)} \tag{23}$$

The term p_i^* is the conditional probability that unit i with observed value y_i belongs to the first component of the mixture model.

Defining $p_{1i}^* = p_i^*$ and $p_{2i}^* = 1 - p_i^*$ for $i = 1, ..., N$, the expectation of the complete-data log-likelihood function $l_C(\vartheta)$ conditional on the observed data is then given by

$$E\left(l_C(\vartheta) | y\right) = \sum_{i=1}^{N} p_{1i}^* \log \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \sum_{i=1}^{N} p_{2i}^* \log \phi_i\left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right) + \sum_{i=1}^{N}\left(p_{1i}^* \log p_i + p_{2i}^* \log\left(1-p_i\right)\right)$$

M-Step: In this step an estimate of ϑ is obtained by maximizing $E\left(l_C(\vartheta) | y\right)$ with respect to ϑ keeping p_{1i}^* and p_{2i}^* fixed. Using the expression above we see that this maximization problem can be

divided into three separate maximization problems that are all well-known. The first two are similar and involve maximizing with respect to the parameters in the component densities. The function to be maximized is $\sum_{i=1}^{N} p_{ji}^{*} \log \phi_i \left(\xi_j, \sigma_{j\varepsilon}^2, q_j\right)$ which can be considered as a weighted maximum likelihood estimation problem involving sums of logarithms weighted by the probabilities that the observations belong to the appropriate component population. This maximization problem is well-known as the component density $\phi_i \left(\xi_j, \sigma_{j\varepsilon}^2, q_j\right)$ given in (13) corresponds to that of a standard random effects model. Obviously, this log-likelihood function can be concentrated with respect to $\sigma_{j\varepsilon}^2$ and the estimator $\hat{\sigma}_{j\varepsilon}^2$ is given by

$$\hat{\sigma}_{j\varepsilon}^2 \left(\xi_j, q_j\right) = \frac{\sum_{i=1}^{N} p_{ji}^{*} \left(y_i - Z_i \xi_j\right)' \left(\left(1+Tq_j\right)^{-1} J_T + C_T\right) \left(y_i - Z_i \xi_j\right)}{T \sum_{i=1}^{N} p_{ji}^{*}} \tag{24}$$

The concentrated log-likelihood function is then maximized with respect to ξ_j and q_j. As there is no explicit solution to this problem it must be solved numerically. The estimators $\hat{\xi}_j$ and $\hat{q}_j$ are obtained by maximizing the following expression with respect to ξ_j and q_j numerically

$$-\sum_{i=1}^{N} p_{ji}^{*} \frac{1}{2} \left(T \log \left(\frac{\sum_{i=1}^{N} p_{ji}^{*} \left(y_i - Z_i \xi_j\right)' \left(\left(1+Tq_j\right)^{-1} J_T + C_T\right) \left(y_i - Z_i \xi_j\right)}{T \sum_{i=1}^{N} p_{ji}^{*}} \right) + \log\left(1+Tq_j\right) \right) \tag{25}$$

This gives the following expression for the estimator $\hat{\sigma}_{j\varepsilon}^2$

$$\hat{\sigma}_{j\varepsilon}^2 \left(\hat{\xi}_j, \hat{q}_j\right) = \frac{\sum_{i=1}^{N} p_{ji}^{*} \left(y_i - Z_i \hat{\xi}_j\right)' \left(\left(1+T\hat{q}_j\right)^{-1} J_T + C_T\right) \left(y_i - Z_i \hat{\xi}_j\right)}{T \sum_{i=1}^{N} p_{ji}^{*}} \tag{26}$$

In the case where $q_j = 0$ it is possible to obtain a closed-form expression for the estimator $\hat{\xi}_j$ which has a weighted form.

The third maximization problem is that of maximizing with respect to γ which is the vector of parameters in the mixing weights. The function to be maximized is $\sum_{i=1}^{N} \left(p_i^{*} \log p_i + \left(1-p_i^{*}\right) \log\left(1-p_i\right)\right)$ which can be considered as a log-likelihood function in a logistic regression model with the pseudo observations p_i^{*}. This means that the estimate of γ is obtained by a logistic regression of p_i^{*} on the variable D_i. As there is no explicit solution of these likelihood equations it is necessary to use a numerical method to obtain estimate γ. The estimator $\hat{\gamma}$ is obtained by using the Newton-Raphson method. Starting with the initial value $\gamma^{(0)}$ the iterations are the following

$$\gamma^{(j+1)} = \gamma^{(j)} + \left(\sum_{i=1}^{N} p_i^{(j)} \left(1 - p_i^{(j)}\right) D_i D_i' \right)^{-1} \left(\sum_{i=1}^{N} \left(p_i^{*} - p_i^{(j)}\right) D_i \right) \tag{27}$$

for $j = 0, 1, 2, ...$ where $p_i^{(j)} = \frac{\exp\left(\gamma^{(j)\prime} D_i\right)}{1+\exp\left(\gamma^{(j)\prime} D_i\right)}$. When the mixing weights are the same for all cross-section units, i.e. $p_i = p = \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)}$ for $i = 1, ..., N$ then

$$\hat{\gamma}_1 = \log\left(\frac{\sum_{i=1}^{N} p_i^{*}}{N - \sum_{i=1}^{N} p_i^{*}} \right) \quad \text{or} \quad \hat{p} = \frac{1}{N} \sum_{i=1}^{N} p_i^{*} \tag{28}$$

The asymptotic variance matrix of the estimator $\hat{\vartheta}$ can be estimated consistently by the observed information matrix. Expressions for the first and second order derivatives of the log-likelihood function defined in (15) are found in Appendix A.1 and A.2.

4 Unit root testing

As explained in the previous sections, the mixture model can be interpreted as resulting from mixing two groups of cross-section units where each group is characterized by the value of the parameters describing the time-series processes for y_{it}. In particular, the value of the AR coefficient can be different in the two groups. So the model provides a natural framework for testing the hypothesis that a group of cross-section units has time-series processes with a unit root.

When no additional restrictions are imposed on the parameters in the model, the time-series processes for y_{it} lead to very different behavior over time depending on whether the AR coefficient is unity or the AR coefficient is less than unity in absolute value. The reason is the presence of the individual-specific terms and the exogenous variables. With respect to the individual-specific terms, they are meant to reflect that different cross-section units have different levels. However, if they are kept unrestricted when the AR coefficient equals unity, they reflect that different cross-section units have different linear time trends. Hence, the individual-specific terms are restricted to zero under the null hypothesis of a unit root. Taking the cross-section dimension of the panel into account, this means that under the null hypothesis all persistency in the time-series processes is attributed to the AR coefficient which is common for all cross-section units in a particular group. On the other hand, under the alternative hypothesis when the AR coefficient is less than unity in absolute value, the persistency in the time-series processes is attributed to both the AR coefficient and the individual-specific term. With respect to exogenous variables, such as time-constant variables and linear time trends, they must appear in such a way that by imposing restrictions on the parameters they lead to similar behavior over time irrespective of the value of the AR coefficient. The issue of how to include deterministic and exogenous variables has been investigated thoroughly in the literature on time-series analysis of non-stationary variables. In this case, the issue is important not only with respect to interpretation of the model but also with respect to the inference procedure, see for example Nielsen & Rahbek (2000). In the present paper, it is only important with respect to interpretation of the model.

More specifically, in a model allowing for a constant and a linear time trend we include the following three $T \times 1$ vectors: $\iota_T = (1, ..., 1)'$, $\tilde{\iota}_T = (0, 1, ..., 1)'$ and $\tau_T = (1, 2, ..., T)'$. Letting x_{it} denote the remaining exogenous variables, we also include the lagged variable x_{it-1} where $x_{i0} \equiv 0$. In this case, the matrix of regressors is $Z_i = \left[\begin{array}{ccccccc} y_{i,-1} & \iota_T & \tilde{\iota}_T & \tau_T & x_i & x_{i,-1} & y_{i0}\iota_T \end{array} \right]$ with corresponding parameter vector $\xi_j = \left(\rho_j, \mu_j, \tilde{\mu}_j, \delta_j, \psi_j, \tilde{\psi}_j, \alpha_j\right)$ for $j = 1, 2$. According to the discussion above, the unit root hypothesis for the cross-section units in group 2 is formulated as

$$H_{01} : \xi_2 = (1, \mu_2, \tilde{\mu}_2, 0, \psi_2, -\psi_2, 0) \text{ and } q_2 = 0 \tag{29}$$

where the parameters $\mu_2, \tilde{\mu}_2$ and ψ_2 are unrestricted. The hypothesis is tested by using a likelihood ratio test statistic which is asymptotically distributed as $\chi^2 (\dim (x_{it}) + 4)$. If this hypothesis is accepted we can go one step further and test the hypothesis

$$H_{02} : \xi_1 = (1, \mu_1, \tilde{\mu}_1, 0, \psi_1, -\psi_1, 0) \text{ and } q_1 = 0 \tag{30}$$

where the parameters $\mu_1, \tilde{\mu}_1$ and ψ_1 are unrestricted. Note that these likelihood ratio test statistics are asymptotically χ^2-distributed when the two components in the mixture model are different. If this is not the case, the usual asymptotic representations are not valid, see the discussion in Section 3.

We define $\Delta y_i = y_i - y_{i,-1}$, $Z_i^R = \left[\begin{array}{ccc} \iota_T & \tilde{\iota}_T & x_i - x_{i,-1} \end{array} \right]$ and the corresponding vector of parameters $\xi^R = \left(\mu, \tilde{\mu}, \psi'\right)$ which is unrestricted. Under the null hypothesis H_{01} the density function is the following

$$f_{\vartheta^R}\left(y_i | \, y_{i0}, D_i, z_i\right) \propto p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1 - p_i\right) \phi_i\left(\xi_2^R, \sigma_{2\varepsilon}^2, 0\right) \tag{31}$$

where as before the unrestricted component density is

$$\phi_i\left(\xi, \sigma_\varepsilon^2, q\right) = \left(\sigma_\varepsilon^2\right)^{-\frac{T}{2}} \left(1 + Tq\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\sigma_\varepsilon^2\right)^{-1} \left(y_i - Z_i\xi\right)' \left(\left(1 + Tq\right)^{-1} J_T + C_T\right) \left(y_i - Z_i\xi\right)\right\} \tag{32}$$

and the restricted component density is

$$\phi_i\left(\xi_2^R, \sigma_{2\varepsilon}^2, 0\right) = \left(\sigma_{2\varepsilon}^2\right)^{-\frac{T}{2}} \exp\left\{-\left(2\sigma_{2\varepsilon}^2\right)^{-1} \left(\Delta y_i - Z_i^R \xi^R\right)' \left(\Delta y_i - Z_i^R \xi^R\right)\right\} \tag{33}$$

As explained in Section 3.2, the maximum likelihood estimates of the parameters are obtained by using the EM algorithm. In the restricted model the E-step and the M-step are as described below.

E-step: The conditional probability that y_i belongs to the first component of the mixture model is calculated as

$$p_i^* = \frac{p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right)}{p_i \phi_i\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) + \left(1 - p_i\right) \phi_i\left(\xi_2^R, \sigma_{1\varepsilon}^2, 0\right)} \tag{34}$$

M-step: The parameters belonging to the first component density are estimated numerically as in the M-step described in Section 3.2. The parameters belonging to the second component density are the closed form expressions given below. With $p_{2i}^* = 1 - p_i^*$ they are on the following forms

$$\hat{\xi}_2^R = \left(\sum_{i=1}^N p_{2i}^* Z_i^{R\prime} Z_i^R\right)^{-1} \sum_{i=1}^N p_{2i}^* Z_i^{R\prime} \Delta y_i \tag{35}$$

$$\hat{\sigma}_{2\varepsilon}^2 = \frac{\sum_{i=1}^N p_{2i}^* \left(\Delta y_i - Z_i^R \hat{\xi}_2^R\right)' \left(\Delta y_i - Z_i^R \hat{\xi}_2^R\right)}{T \sum_{i=1}^N p_{2i}^*} \tag{36}$$

As before the parameter γ describing the mixing weights is estimated by a logistic regression of p_i^* on the variable D_i.

5 Empirical application

The data set is drawn from the Panel Study of Income Dynamics (PSID). The PSID is a panel data set beginning in 1968 with approximately 4,800 families. Of these almost 40% are low-income families from the Survey of Economic Opportunity (SEO). With respect to income the SEO subsample is not random whereas the remaining families can be considered as a random sample of US families. The survey follows individuals that are members of the families drawn in 1968 as well as their offspring and individuals entering the families for example by marriage.

We use a data set covering the period 1969-93. The data set consists of males aged 25-55 who are head of the household, report positive earnings and do not belong to the SEO. In addition we only include individuals with a constant level of education. All individuals that are observed for at least 15 adjoining years and satisfy the criteria described above in every year (in which they are in the PSID) are included in our data set. The earnings variable is defined as annual income from labor and corresponds to the year before the interview. The annual earnings are deflated by the CPI (base 1982-84). From this data set we eliminate individuals where the log of real earnings is less than 8.5 or greater than 12.5 in at least one year. This means that sample is very homogenous with respect to earnings. Altogether, we end up with an unbalanced panel data set which includes observations from 562 individuals and a total of 10,890 individual-year observations. Table 4 shows the number of individuals observed in each year, and the mean and standard deviation of log earnings in each year. We see that all individuals are observed in the period 1978-82. Finally, the individuals are divided into the following three education groups: high school dropouts (individuals with less than 12 grades of schooling), high school graduates (individuals with at least a high school diploma but no college degree) and college graduates (individuals with a college degree or more). These education groups correspond to respectively 10.1%, 68.2% and 21.7% of the individuals.

We assume that the log of annual earnings are generated by the mixture model defined in Section 2. As exogenous variables we include: a constant, a linear time trend, dummies for high school dropouts and college graduates, age and age squared. As explained in Section 4 the lagged values of these variables are also included. This means that the number of linear independent columns in Z_i is 6. As each individual is observed for at least 15 years, i.e. $T \geq 14$ for all individuals, the restriction in Theorem 1 is satisfied. We have tried to include the education dummies as explanatory variables in the mixing weights, see the estimation results reported in Table 5 in Appendix B. However, these variables do not have a significant effect on the mixing weights. Therefore the mixing weights are assumed to be the same for all individuals, i.e. $p_i = p$ for $i = 1, ..., N$. The parameter estimates from this model are reported in the Table 1.

The estimates of the AR coefficients are 0.50 and 0.71. So none of them are close to unity. The estimate of the probability that a given individual belongs to group 1 is 0.49, so the proportions of the two groups are equal. In addition, the estimate of the error variance σ_ε^2 is around 10 times higher in group 1 compared to group 2 and the estimate of the individual-specific variation σ_α^2 is around 20% of the error variance σ_ε^2 in both groups. Altogether, the short-run variation is much higher in group 1 than in group 2. We will return to the estimates of the parameters describing the level of log earnings below.

The parameter estimates from the model where the unit root hypothesis H_{01} defined in (29) is imposed are reported in Table 2. The unit root hypothesis is clearly rejected - a $\chi^2(7)$ of 973.06. This finding is in contrast to the papers by McCurdy (1982) and Abowd & Card (1989) where the autoregressive coefficient as a starting point is assumed to be unity. Our analysis does not support this assumption.

Table 1: Parameter estimates from the mixture model (standard errors are in brackets)

Variable	Group 1	Group 2
Lagged log earnings ρ	0.5043 (0.0145)*	0.7060 (0.0150)*
Constant	2.3185 (0.9674)*	0.5709 (0.2582)*
Dummy for high school dropout	-0.1780 (0.0744)*	-0.0167 (0.0323)
Dummy for college graduate	0.0659 (0.0540)	0.0657 (0.0232)*
Linear time trend	-0.0003 (0.0021)	0.0021 (0.0009)*
Age	0.0446 (0.0579)	0.0276 (0.0133)*
Age^2/100	-0.0561 (0.0852)	-0.0432 (0.0194)*
Lagged constant	0.0588 (0.9250)	0.2561 (0.2111)
Lagged dummy for high school dropout	0.0989 (0.0699)	-0.0225 (0.0295)
Lagged dummy for college graduate	0.0898 (0.0509)	0.0195 (0.0210)
Lagged age	-0.0045 (0.0579)	-0.0203 (0.0130)
Lagged age^2/100	0.0065 (0.0873)	0.0349 (0.0195)
Initial log earnings α	0.1853 (0.0240)*	0.2016 (0.0174)*
Short-run variance σ_ε^2	0.1074 (0.0025)*	0.0170 (0.0005)*
$q = \sigma_\alpha^2/\sigma_\varepsilon^2$	0.1885 (0.0263)*	0.2593 (0.0406)*
Mixing weight p	0.4885 (0.0230)*	0.5115 (0.0230) *
Log-likelihood: 684.69		

* The parameter is significantly different from zero at the 5% level

Table 2: Parameter estimates from the mixture model under the unit root hypothesis H_{01} (standard errors are in brackets)

Variable	Group 1	Group 2
Lagged log earnings ρ	0.6616 (0.0169)*	1
Constant	0.6189 (0.2671)*	-0.9122 (0.4306)*
Dummy for high school dropout	-0.0381 (0.0311)	0.9265 (0.4326)*
Dummy for college graduate	0.0767 (0.0244)*	-0.1407 (0.1001)
Linear time trend	0.0024 (0.0010)*	0
Age	0.0466 (0.0154)*	0.0594 (0.0255)*
Age^2/100	-0.0687 (0.0222)*	-0.0887 (0.0372)*
Lagged constant	0.5519 (0.2427)*	0.0075 (0.0664)
Lagged dummy for high school dropout	0.0022 (0.0277)	-0.9265
Lagged dummy for college graduate	0.0220 (0.0225)	0.1407
Lagged age	-0.0375 (0.0149)*	-0.0594
Lagged age^2/100	0.0586 (0.0221)*	0.0887
Initial log earnings α	0.2075 (0.0166)*	0
Short-run variance σ_ε^2	0.0212 (0.0011)*	0.1608 (0.0060)*
$q = \sigma_\alpha^2/\sigma_\varepsilon^2$	0.3035 (0.0430)*	0
Mixing weight p	0.5795 (0.0292)*	0.4205 (0.0292) *
Log-likelihood: 198.16		

* The parameter is significantly different from zero at the 5% level

In Table 1, the parameters corresponding to the lagged variables do not seem to be significantly different from zero. Therefore, we estimate the model when only a constant, age, age squared and education dummies are included as exogenous variables. The parameter estimates from this model are reported in Table 3.

Table 3: Parameter estimates from the mixture model (standard errors are in brackets)

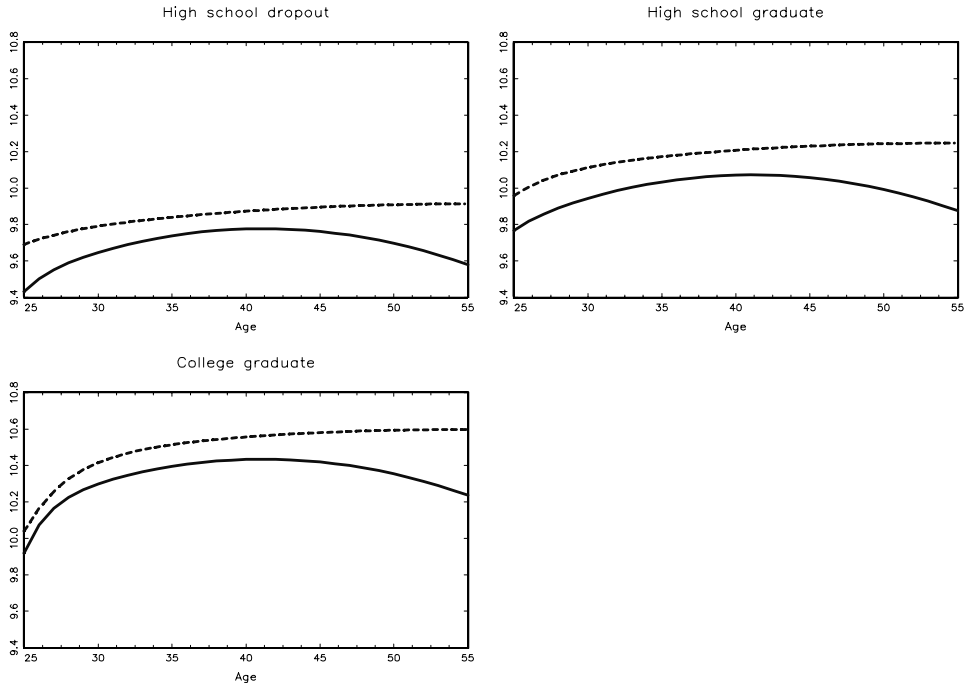
Variable	Group 1	Group 2
Lagged log earnings ρ	0.5051 (0.0145)*	0.7099 (0.0150)*
Constant	2.3629 (0.2539)*	0.9452 (0.1447)*
Dummy for high school dropout	-0.0846 (0.0338)*	-0.0465 (0.0153)*
Dummy for college graduate	0.1509 (0.0246)*	0.0863 (0.0120)*
Age	0.0409 (0.0063)*	0.0056 (0.0027)*
$\text{Age}^2/100$	-0.0510 (0.0080)*	-0.0054 (0.0033)
Initial log earnings α	0.1845 (0.0229)*	0.1889 (0.0156)*
Short-run variance σ_ε^2	0.1074 (0.0025)*	0.0171 (0.0005)*
$q = \sigma_\alpha^2/\sigma_\varepsilon^2$	0.1865 (0.0261)*	0.2577 (0.0409)*
Mixing weight p	0.4883 (0.0232)*	
Log-likelihood: 672.86		

* The parameter is significantly different from zero at the 5% level

As described above, the estimate of the AR coefficient is higher and the estimate of the variance is lower in group 2 compared to group 1. In addition, the estimates of the constant for all education groups is higher in group 1 compared to group 2 and the estimates of the parameter α describing the conditional mean effect of the initial values on the individual-specific terms are equal in the two groups. The age effect in group 1 is hump-shaped with a peak at the age of 40 while the age effect in group 2 is very low and linear in age. In Figure 1 we show the mean level of log earnings as a function of age for two individuals from different groups. The initial log earnings at the age of 25 is calculated as the weighted average of log earnings for individuals at the age 25 where p_i^* and $(1 - p_i^*)$ are used as weights. We see that for all education groups the average level is higher in group 2 compared to group 1. For an individual at age 30 the level of annual earnings in group 2 is around 3% higher compared to group 1. For individuals at age 50 the number is 6%. In addition, the variation group 1 is much higher than in group 2.

Finally, in their analysis of income data from the PSID, Alvarez, Browning & Ejrnæs (2002) estimate the mean AR coefficient as being 0.85 whereas we estimate it as being 0.61. The difference is probably explained by the fact that our models are different in an important aspect. In Alvarez, Browning & Ejrnæs (2002) the individual-specific effects are assumed to be proportional to the initial values without any additional variation. In our model, this corresponds to $q_1 = q_2 = 0$. For comparison, our model is estimated when this restriction is imposed. These parameter estimates are reported in Table 6 in Appendix B. We see that this increases the estimates of the AR coefficients to 0.67 and 0.88 with an estimate of the mean AR coefficient which is 0.78. This is quite close to the estimate in Alvarez,

Figure 1: Mean log earnings for group 1 (solid line) and group 2 (dashed line)



Browning & Ejrnæs (2002). The hypothesis that $q_1 = q_2 = 0$ is clearly rejected in our sample - a $\chi^2(2)$ of 321.06 - so it seems to be important to allow for additional variation in the individual-specific effects.

6 Conclusions

In this paper we have considered estimation of the parameters in a dynamic panel data model where all parameters including the regression coefficients differ across cross-section units. To do this we have used an approach where the parameter variation is assumed to be discrete. The model can be interpreted as resulting from mixing two groups of cross-section units where each group is characterized by the value of the parameters in their time-series processes. An important feature of the model is that it allows for exogenous variables and heterogeneity through these.

We have shown that when the number of exogenous variables is restricted in relation to the time-series dimension of the panel, the maximum likelihood estimator of the parameters in the model has the usual asymptotic properties. In particular, the singularity problem usually encountered in this type of mixture model can be avoided. The model provides a framework for testing unit root hypotheses that can not be tested by existing test procedures. More specifically, it is possible to test the hypothesis that a group of cross-section units has time-series processes with a unit root. If the hypothesis is accepted, the proportion of this group and information about which cross-section units are most likely to belong to this group is automatically provided.

The method is applied to income data on individuals drawn from the PSID. In this sample there is

no evidence of unit roots. However, it is clear that the income processes in the two groups of individuals are very different. More specifically, we find that there is a negative correlation between the level and the variation of income. This means that individuals in the group with a high average level of income also have low variation in their income and vice versa. The mixing proportions of the two groups is the same for all individuals.

Finally, there are some issues that are left for future research. First, it is important to verify that the model provides an adequate fit of the data by performing some specification tests. Adapting the approach suggested by Bhargava & Sargan (1983), this can be done by testing the validity of the restrictions being imposed on the time-series behavior of the regression errors. More specifically, by testing the null hypothesis that the time-series behavior of the regression errors is on the same form as in a standard random effects models. The hypothesis is tested against alternatives where the time-series behavior of the regression errors is on a more general form, such as a completely unrestricted form or forms with systematic autocorrelation. The issue of misspecification testing is related to the problem of how to determine the number of components in the mixture model. Clearly, the mixture model considered in this paper can easily be extended to allow for a finite number of components instead of just two. In this case, the problem of how to determine the number of components arises. If the model does not allow for sufficient variation in the regression coefficients, this is likely to give serial correlation in the regression errors. This means that the time-series behavior of the regression errors will be different from that in a random effects model. Hence, if the model seems misspecified, this could be an indication that there is additional variation in the regression coefficients. In this case, an extra component is included in the mixture model and the specification test is performed again.

Second, when we find that the autoregressive coefficients are less than unity in absolute value, it might be of interest to test the hypothesis that the parameters describing the long-run relations between the variables are the same for all cross-section units. In dynamic macro-panels where the cross-section and the time-series dimensions are similar in magnitude, estimation of the parameters when this restriction is imposed as a starting point has been investigated by Pesaran, Shin & Smith (1999). The reason for the interest in this hypothesis is that usually economic theory is concerned with long-run relations between variables. Hence, there might be good reasons to expect these to be the same for all cross-section units.

A Appendix

Throughout this appendix the following notation will be used

$$\phi_i^j = \phi_i\left(\xi_j, \sigma_{j\varepsilon}^2, q_j\right) \quad \text{for } j = 1, 2 \tag{37}$$

$$r_i = \frac{p_i}{1-p_i} = \exp\left(\gamma' D_i\right) \tag{38}$$

$$p_i^* = \frac{p_i \phi_i^1}{p_i \phi_i^1 + (1-p_i)\,\phi_i^2} \tag{39}$$

The well-known results provided in Lemma 1 below are useful.

Lemma 1 *Let ε be a $n \times 1$ vector where $\varepsilon \sim N(0, \Omega)$ and let A, B be $n \times n$ matrices. Then*

$$E(A\varepsilon) = 0 \tag{40}$$

$$E(\varepsilon' A\varepsilon) = \operatorname{tr}\{A\Omega\} \tag{41}$$

$$E(\varepsilon\varepsilon' A\varepsilon) = 0 \tag{42}$$

$$E(\varepsilon' A\varepsilon\varepsilon' B\varepsilon) = \operatorname{tr}\{A\Omega B\Omega\} + \operatorname{tr}\{A'\Omega B\Omega\} + \operatorname{tr}\{A\Omega\}\operatorname{tr}\{B\Omega\} \tag{43}$$

A.1 First order derivatives of the log-likelihood function

Using the expression for the component density $\phi\left(\xi, \sigma_\varepsilon^2, q\right)$ in (13) we obtain the following expressions for the first order derivatives of $\log \phi_i$ with respect to $\xi, \sigma_\varepsilon^2$ and q

$$\frac{\partial \log \phi_i}{\partial \xi} = Z_i' V^{-1}(y_i - Z_i\xi) \tag{44}$$

$$\frac{\partial \log \phi_i}{\partial \sigma_\varepsilon^2} = \frac{1}{2\sigma_\varepsilon^2}\left(-T + (y_i - Z_i\xi)' V^{-1}(y_i - Z_i\xi)\right) \tag{45}$$

$$\frac{\partial \log \phi_i}{\partial q} = \frac{T}{2(1+Tq)}\left(-1 + \frac{1}{\sigma_\varepsilon^2(1+Tq)}(y_i - Z_i\xi)' J_T (y_i - Z_i\xi)\right) \tag{46}$$

The first order derivatives of the log-likelihood function defined in equation (14) can be expressed as

$$\frac{\partial l_i}{\partial \vartheta_k} = p_i^* \frac{\partial \log \phi_i^1}{\partial \vartheta_k} + (1 - p_i^*)\frac{\partial \log \phi_i^2}{\partial \vartheta_k} + (p_i^* - p_i)\frac{\partial \log r_i}{\partial \vartheta_k} \tag{47}$$

where ϑ_k is an element in $\vartheta = \left(\xi_1', \sigma_{1\varepsilon}^2, q_1, \xi_2', \sigma_{2\varepsilon}^2, q_2, \gamma'\right)'$.

Using this yields the following

$$\frac{\partial l_i(\vartheta)}{\partial \xi_1} = p_i^* Z_i' V_1^{-1}(y_i - Z_i\xi_1) \tag{48}$$

$$\frac{\partial l_i(\vartheta)}{\partial \sigma_{1\varepsilon}^2} = p_i^* \frac{1}{2\sigma_{1\varepsilon}^2}\left(-T + (y_i - Z_i\xi_1)' V_1^{-1}(y_i - Z_i\xi_1)\right) \tag{49}$$

$$\frac{\partial l_i(\vartheta)}{\partial q_1} = p_i^* \frac{T}{2(1+Tq_1)}\left(-1 + \frac{1}{\sigma_{1\varepsilon}^2(1+Tq_1)}(y_i - Z_i\xi_1)' J_T(y_i - Z_i\xi_1)\right) \tag{50}$$

$$\frac{\partial l_i(\vartheta)}{\partial \xi_2} = (1-p_i^*) Z_i' V_2^{-1}(y_i - Z_i\xi_2) \tag{51}$$

$$\frac{\partial l_i(\vartheta)}{\partial \sigma_{2\varepsilon}^2} = (1-p_i^*) \frac{1}{2\sigma_{2\varepsilon}^2}\left(-T + (y_i - Z_i\xi_2)' V_2^{-1}(y_i - Z_i\xi_2)\right) \tag{52}$$

$$\frac{\partial l_i(\vartheta)}{\partial q_2} = (1-p_i^*) \frac{T}{2(1+Tq_2)}\left(-1 + \frac{1}{\sigma_{2\varepsilon}^2(1+Tq_2)}(y_i - Z_i\xi_2)' J_T(y_i - Z_i\xi_2)\right) \tag{53}$$

$$\frac{\partial l_i(\vartheta)}{\partial \gamma} = (p_i^* - p_i) D_i \tag{54}$$

A.2 Second order derivatives of the log-likelihood function

Using the expression for the component density $\phi\left(\xi, \sigma_\varepsilon^2, q\right)$ in (13) we obtain the following expressions for the second order derivatives of $\log \phi_i$ with respect to $\xi, \sigma_\varepsilon^2$ and q

$$\frac{\partial^2 \log \phi_i}{\partial \xi \partial \xi'} = -Z_i' V^{-1} Z_i \tag{55}$$

$$\frac{\partial^2 \log \phi_i}{(\partial \sigma_\varepsilon^2)^2} = \frac{1}{2\sigma_\varepsilon^4}\left(T - 2(y_i - Z_i\xi)' V^{-1}(y_i - Z_i\xi)\right) \tag{56}$$

$$\frac{\partial^2 \log \phi_i}{(\partial q)^2} = \frac{T^2}{2(1+Tq)^2}\left(1 - \frac{2}{\sigma_\varepsilon^2(1+Tq)}(y_i - Z_i\xi)' J_T(y_i - Z_i\xi)\right) \tag{57}$$

$$\frac{\partial^2 \log \phi_i}{\partial \xi \partial \sigma_\varepsilon^2} = -\frac{1}{2\sigma_\varepsilon^2} Z_i' V^{-1}(y_i - Z_i\xi) \tag{58}$$

$$\frac{\partial^2 \log \phi_i}{\partial \xi \partial q} = -\frac{T}{\sigma_\varepsilon^2(1+Tq)^2} Z_i' J_T(y_i - Z_i\xi) \tag{59}$$

$$\frac{\partial^2 \log \phi_i}{\partial \sigma_\varepsilon^2 \partial q} = -\frac{T}{2\sigma_\varepsilon^4(1+Tq)^2}(y_i - Z_i\xi)' J_T(y_i - Z_i\xi) \tag{60}$$

Using the expression in (47) and that all second order derivatives of $\log r_i = \gamma' D_i$ are equal to zero, the second order derivatives of the log-likelihood function defined in equation (14) can be expressed as

$$\frac{\partial^2 l_i}{\partial \vartheta_k \partial \vartheta_l} = \frac{\partial p_i^*}{\partial \vartheta_l}\frac{\partial \log \phi_i^1}{\partial \vartheta_k} + p_i^* \frac{\partial^2 \log \phi_i^1}{\partial \vartheta_k \partial \vartheta_l} - \frac{\partial p_i^*}{\partial \vartheta_l}\frac{\partial \log \phi_i^2}{\partial \vartheta_k} + (1-p_i^*)\frac{\partial^2 \log \phi_i^2}{\partial \vartheta_k \partial \vartheta_l} + \left(\frac{\partial p_i^*}{\partial \vartheta_l} - \frac{\partial p_i}{\partial \vartheta_l}\right)\frac{\partial \log r_i}{\partial \vartheta_k}$$

where ϑ_k, ϑ_l are elements in ϑ. By inserting

$$\frac{\partial p_i}{\partial \vartheta_k} = p_i(1-p_i)\frac{\partial \log r_i}{\partial \vartheta_k} \tag{61}$$

$$\frac{\partial p_i^*}{\partial \vartheta_k} = p_i^*(1-p_i^*)\left(\frac{\partial \log \phi_i^1}{\partial \vartheta_k} - \frac{\partial \log \phi_i^2}{\partial \vartheta_k} + \frac{\partial \log r_i}{\partial \vartheta_k}\right) \tag{62}$$

we obtain the following expression

$$\begin{aligned}\frac{\partial^2 l_i}{\partial \vartheta_k \partial \vartheta_l} = {} & p_i^* \frac{\partial^2 \log \phi_i^1}{\partial \vartheta_k \partial \vartheta_l} + (1-p_i^*)\frac{\partial^2 \log \phi_i^2}{\partial \vartheta_k \partial \vartheta_l} - p_i(1-p_i)\frac{\partial \log r_i}{\partial \vartheta_k}\frac{\partial \log r_i}{\partial \vartheta_l} \\ & + p_i^*(1-p_i^*)\left(\frac{\partial \log \phi_i^1}{\partial \vartheta_k} - \frac{\partial \log \phi_i^2}{\partial \vartheta_k} + \frac{\partial \log r_i}{\partial \vartheta_k}\right)\left(\frac{\partial \log \phi_i^1}{\partial \vartheta_l} - \frac{\partial \log \phi_i^2}{\partial \vartheta_l} + \frac{\partial \log r_i}{\partial \vartheta_l}\right)\end{aligned} \tag{63}$$

Using this in combination with the expression in (47) gives

$$\begin{aligned}
&\frac{\partial^2 l_i}{\partial\vartheta_k\partial\vartheta_l}+\frac{\partial l_i}{\partial\vartheta_k}\frac{\partial l_i}{\partial\vartheta_l}\\
&=p_i^*\left(\frac{\partial^2\log\phi_i^1}{\partial\vartheta_k\partial\vartheta_l}+\frac{\partial\log\phi_i^1}{\partial\vartheta_k}\frac{\partial\log\phi_i^1}{\partial\vartheta_l}\right)+(1-p_i^*)\left(\frac{\partial^2\log\phi_i^2}{\partial\vartheta_k\partial\vartheta_l}+\frac{\partial\log\phi_i^2}{\partial\vartheta_k}\frac{\partial\log\phi_i^2}{\partial\vartheta_l}\right)\\
&+(p_i^*-p_i)(1-2p_i)\frac{\partial\log r_i}{\partial\vartheta_k}\frac{\partial\log r_i}{\partial\vartheta_l}+p_i^*(1-p_i)\frac{\partial\log\phi_i^1}{\partial\vartheta_k}\frac{\partial\log r_i}{\partial\vartheta_l}-(1-p_i^*)p_i\frac{\partial\log\phi_i^2}{\partial\vartheta_k}\frac{\partial\log r_i}{\partial\vartheta_l}
\end{aligned}\tag{64}$$

Inserting the expressions for the first and second order derivatives of $\log\phi_i^1$ and $\log\phi_i^2$ given above in (44)-(46) and (55)-(60) we have

$$\frac{\partial^2 l_i}{\partial\vartheta\partial\vartheta'}+\frac{\partial l_i}{\partial\vartheta}\frac{\partial l_i}{\partial\vartheta'}=$$

$$\begin{bmatrix}
p_i^*H^{\xi_1\xi_1} & & & & & & \\
p_i^*H^{\sigma_{1\varepsilon}^2\xi_1} & p_i^*H^{\sigma_{1\varepsilon}^2\sigma_{1\varepsilon}^2} & & & & & \\
p_i^*H^{q_1\xi_1} & p_i^*H^{q_1\sigma_{1\varepsilon}^2} & p_i^*H^{q_1q_1} & & & & \\
0 & 0 & 0 & (1-p_i^*)H^{\xi_2\xi_2} & & & \\
0 & 0 & 0 & (1-p_i^*)H^{\sigma_{2\varepsilon}^2\xi_2} & (1-p_i^*)H^{\sigma_{2\varepsilon}^2\sigma_{2\varepsilon}^2} & & \\
0 & 0 & 0 & (1-p_i^*)H^{q_2\xi_2} & (1-p_i^*)H^{q_2\sigma_{2\varepsilon}^2} & (1-p_i^*)H^{q_2q_2} & \\
p_i^*H^{\gamma\xi_1} & p_i^*H^{\gamma\sigma_{1\varepsilon}^2} & p_i^*H^{\gamma q_1} & (1-p_i^*)H^{\gamma\xi_2} & (1-p_i^*)H^{\gamma\sigma_{2\varepsilon}^2} & (1-p_i^*)H^{\gamma q_2} & H^{\gamma\gamma}
\end{bmatrix}$$

where for $j=1,2$

$$H^{\xi_j\xi_j} = -Z_i'V_j^{-1}Z_i+Z_i'V_j^{-1}e_{ji}e_{ji}'V_j^{-1}Z_i \tag{65}$$

$$H^{\sigma_{j\varepsilon}^2\sigma_{j\varepsilon}^2} = \frac{1}{4\sigma_{j\varepsilon}^4}\left(\left(-T+e_{ji}'V_j^{-1}e_{ji}\right)^2+2\left(T-2e_{ji}'V_j^{-1}e_{ji}\right)\right) \tag{66}$$

$$H^{q_jq_j} = \frac{T^2}{4(1+Tq_j)^2}\left(\left(-1+\frac{1}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_Te_{ji}\right)^2+2\left(1-\frac{2}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_Te_{ji}\right)\right) \tag{67}$$

$$H^{\sigma_{j\varepsilon}^2\xi_j} = \frac{1}{2\sigma_{j\varepsilon}^2}\left(e_{ji}'V_j^{-1}e_{ji}-1-T\right)e_{ji}'V_j^{-1}Z_i \tag{68}$$

$$H^{q_j\xi_j} = \frac{T}{2(1+Tq_j)}\left(\left(-1+\frac{1}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_Te_{ji}\right)e_{ji}'V_j^{-1}Z_i-\frac{2}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_TZ_i\right) \tag{69}$$

$$H^{q_j\sigma_{j\varepsilon}^2} = \frac{T}{4\sigma_{j\varepsilon}^2(1+Tq_j)}\left(T-e_{ji}'V_j^{-1}e_{ji}+\frac{1}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_Te_{ji}\left(e_{ji}'V_j^{-1}e_{ji}-2-T\right)\right) \tag{70}$$

$$H^{\gamma\gamma} = (p_i^*-p_i)(1-2p_i)D_iD_i' \tag{71}$$

$$H^{\gamma\xi_j} = (1-p_i)^{2-j}p_i^{j-1}D_ie_{ji}V_j^{-1}Z_i \tag{72}$$

$$H^{\gamma\sigma_{j\varepsilon}^2} = (1-p_i)^{2-j}p_i^{j-1}D_i\left(-T+e_{ji}'V_j^{-1}e_{ji}\right) \tag{73}$$

$$H^{\gamma q_j} = (1-p_i)^{2-j}p_i^{j-1}D_i\frac{T}{2(1+Tq_j)}\left(-1+\frac{1}{\sigma_{j\varepsilon}^2(1+Tq_j)}e_{ji}'J_Te_{ji}\right) \tag{74}$$

with the $T\times 1$ vectors e_{1i} and e_{2i} defined by

$$e_{ji}=y_i-Z_i\xi_j \quad \text{for } j=1,2 \tag{75}$$

A.3 Proof of Theorem 1

The theorem is proved by verifying the standard regularity conditions due to Cramér (1946). The conditions are found in various forms in Section 4 in Amemiya (1985).

Condition 1 *The density function* $f_\vartheta\left(y_i|\, y_{i0}, D_i, z_i\right)$ *satisfies the following conditions:*

(i) *The log-likelihood function* $l_i\left(\vartheta\right) = \log f_\vartheta$ *is well-defined except for a set of measure zero with respect to* $f_\vartheta\left(y_i|\, y_{i0}, D_i, z_i\right)$. *In addition* f_ϑ *is two times continuously differentiable in* ϑ.

(ii) *For all* ϑ *the following hold*

$$E\left(\frac{\partial l_i\left(\vartheta\right)}{\partial\vartheta}\right) = 0 \tag{76}$$

$$E\left\|\frac{\partial l_i\left(\vartheta\right)}{\partial\vartheta}\right\|^2 < \infty \tag{77}$$

$$E\left(\frac{\partial l_i\left(\vartheta\right)}{\partial\vartheta}\frac{\partial l_i\left(\vartheta\right)}{\partial\vartheta'}\right) = -E\left(\frac{\partial^2 l_i\left(\vartheta\right)}{\partial\vartheta\partial\vartheta'}\right) > 0 \tag{78}$$

(iii) *For all* ϑ *there exists a neighborhood* $N\left(\vartheta\right)$ *of* ϑ *such that*

$$E\sup_{\tilde{\vartheta}\in N(\vartheta)}\left|\frac{\partial^3 l_i\left(\tilde{\vartheta}\right)}{\partial\vartheta_k\partial\vartheta_l\partial\vartheta_m}\right| < \infty \tag{79}$$

where $\vartheta_k, \vartheta_l, \vartheta_m$ *are elements of* ϑ.

Lemma 2 *The density function defined in* (12) *satisfies Condition 1.*

Proof of Lemma 2:

First of all we note that by definition

$$p_i^* = E\left[s_i\,|y_{i0}, D_i, z_i, y_i\right] \tag{80}$$

$$p_i = E\left[s_i\,|y_{i0}, D_i, z_i\right] = \Pr\left(s_i = 1\,|y_{i0}, D_i, z_i\right) \tag{81}$$

As explained in Section 3.2, the density function in (12) can be interpreted as coming from the following mixture model

$$\begin{array}{ll} y_i = Z_i\xi_1 + \upsilon_{1i} & \text{if } s_i = 1 \\ y_i = Z_i\xi_2 + \upsilon_{2i} & \text{if } s_i = 0 \end{array} \quad \text{for } i = 1, ..., N \tag{82}$$

where $\left(\upsilon_{1i}|\, y_{i0}, D_i, z_i\right)$ is iid$N\left(0, V_1\right)$ and $\left(\upsilon_{2i}|\, y_{i0}, D_i, z_i\right)$ is iid$N\left(0, V_2\right)$ with $V_j = \sigma_{j\varepsilon}^2 I_T + \sigma_{j\alpha}^2 \iota_T\iota_T'$ for $j = 1, 2$. Using that $Z_i = \left[\begin{array}{ccc} y_{i,-1} & z_i & y_{i0}\iota_T \end{array}\right]$ we can express the model above as

$$\begin{array}{ll} y_{it} = \rho_1 y_{it-1} + z_{it}'\omega_1 + \alpha_1 y_{i0} + \upsilon_{1,it} & \text{for } t = 1, ..., T \quad \text{if } s_i = 1 \\ y_{it} = \rho_2 y_{it-1} + z_{it}'\omega_2 + \alpha_2 y_{i0} + \upsilon_{2,it} & \text{for } t = 1, ..., T \quad \text{if } s_i = 0 \end{array} \quad \text{for } i = 1, ..., N \tag{83}$$

where $\upsilon_{j,it}$ denotes element t in υ_{ji} for $j = 1, 2$. By recursive substitution in the expression above we find that

$$y_{i,-1} = \begin{cases} A\left(\rho_1\right)\upsilon_{1i} + Q\left(y_{i0}, z_i, \xi_1\right) & \text{if } s_i = 1 \\ A\left(\rho_2\right)\upsilon_{2i} + Q\left(y_{i0}, z_i, \xi_2\right) & \text{if } s_i = 0 \end{cases} \tag{84}$$

where $Q(y_{i0}, z_i, \xi)$ is a $T \times 1$ vector where element t for $t = 1, ..., T$ equals $\left(\rho^{t-1} + \left(1 + ... + \rho^{t-2}\right)\alpha\right) y_{i0} + \left(\sum_{s=0}^{t-2} \rho^s z_{it-1-s}\right)' \omega$ and $A(\rho)$ is the following $T \times T$ matrix

$$A(\rho) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & \vdots & \vdots \\ \rho & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \rho^{T-2} & \cdots & \rho & 1 & 0 \end{bmatrix} \tag{85}$$

Using the expression in (84) we can express $Z_i = \begin{bmatrix} y_{i,-1} & z_i & y_{i0}\iota_T \end{bmatrix}$ in the following way

$$Z_i = \begin{cases} Z_{1i} & \text{if } s_i = 1 \\ Z_{2i} & \text{if } s_i = 0 \end{cases} \quad \text{for } Z_{ji} = \begin{bmatrix} A(\rho_j)\upsilon_{ji} + Q(y_{i0}, z_i, \xi_j) & \vdots & R(y_{i0}, z_i) \end{bmatrix} \quad \text{for } j = 1, 2 \tag{86}$$

where $R(y_{i0}, z_i)$ is the $T \times (k+1)$ matrix defined as $R(y_{i0}, z_i) = \begin{bmatrix} z_i & y_{i0}\iota_T \end{bmatrix}$.

The results below are used in the following

$$s_i K(e_{1i}, Z_i) = \begin{cases} K(\upsilon_{1i}, Z_{1i}) & \text{if } s_i = 1 \\ 0 & \text{if } s_i = 0 \end{cases} \tag{87}$$

$$(1 - s_i) K(e_{2i}, Z_i) = \begin{cases} 0 & \text{if } s_i = 1 \\ K(\upsilon_{2i}, Z_{2i}) & \text{if } s_i = 0 \end{cases} \tag{88}$$

where $K(e_{ji}, Z_i)$ denotes a matrix where the elements depend on $e_{ji} = y_i - Z_i\xi_j$ and Z_i for $j = 1, 2$.

In addition the following results hold

$$\operatorname{tr}\left\{A(\rho_j)\right\} = 0 \tag{89}$$

$$\operatorname{tr}\left\{A(\rho_j) A(\rho_j)\right\} = 0 \tag{90}$$

$$\operatorname{tr}\left\{J_T V_j\right\} = \sigma_{j\varepsilon}^2 (1 + Tq_j) \tag{91}$$

The results in (89) and (90) follow by using the expression for $A(\rho)$ in (85). The result in (91) follows by using that $V_j = \sigma_{j\varepsilon}^2 (1 + Tq_j) J_T + \sigma_{j\varepsilon}^2 C_T$ where $J_T = \iota_T \iota_T'/T$ and $C_T = I_T - J_T$ such that J_T is idempotent and $J_T C_T = 0$. Using this we have $J_T V_j = \sigma_{j\varepsilon}^2 (1 + Tq_j) J_T$ such that $\operatorname{tr}\{J_T V_j\} = \sigma_{j\varepsilon}^2 (1 + Tq_j) \operatorname{tr}\{J_T\} = \sigma_{j\varepsilon}^2 (1 + Tq_j)$.

We obtain the following expressions for the conditional means

$$E\left[V_j^{-1}\upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] = 0 \tag{92}$$

$$E\left[\upsilon_{ji}' V_j^{-1} \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] = \operatorname{tr}\{I_T\} = T \tag{93}$$

$$E\left[\upsilon_{1i}' J_T \upsilon_{1i} \middle| y_{i0}, D_i, z_i\right] = \operatorname{tr}\{J_T V_j\} = \sigma_{j\varepsilon}^2 (1 + Tq_j) \tag{94}$$

$$E\left[\upsilon_{ji}' A(\rho_j)' V_j^{-1} \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] = \operatorname{tr}\left\{A(\rho_j)' V_j^{-1} V_j\right\} = \operatorname{tr}\left\{A(\rho_j)\right\} = 0 \tag{95}$$

They follow by Lemma 1 as $(\upsilon_{ji}|\, y_{i0}, D_i, z_i) \sim N(0, V_j)$ and the results in (89)-(91).

(i) Clearly, this condition is satisfied.

(ii) To show (76) we first of all note the following

$$E\left[R(y_{i0}, z_i)' V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] = R(y_{i0}, z_i)' E\left[V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] = 0$$

Also we have

$$\begin{aligned} & E\left[\left(A(\rho_j)\,\upsilon_{ji} + Q(y_{i0}, z_i, \xi_j)\right)' V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] \\ = \; & E\left[\upsilon_{ji}' A(\rho_j)' V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] + Q(y_{i0}, z_i, \xi_j)' E\left[V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] \\ = \; & 0 \end{aligned}$$

Comparing with the expression for Z_{ji} in (86) this shows that

$$E\left[Z_{ji}' V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] = 0 \tag{96}$$

Using the expression in (48) we have

$$\begin{aligned} & E\left[\frac{\partial l_i(\vartheta)}{\partial \xi_1} \middle|\, y_{i0}, D_i, z_i\right] \\ = \; & E\left[E[s_i|\, y_{i0}, D_i, z_i, y_i]\, Z_i' V_1^{-1} e_{1i} \middle|\, y_{i0}, D_i, z_i\right] && (97) \\ = \; & E\left[s_i Z_i' V_1^{-1} e_{1i} \middle|\, y_{i0}, D_i, z_i\right] \\ = \; & p_i E\left[Z_{1i}' V_1^{-1} \upsilon_{1i} \middle|\, y_{i0}, D_i, z_i\right] && (98) \\ = \; & 0 && (99) \end{aligned}$$

where the first equality sign follows by inserting the expression for p_i^* in (80), the second equality sign follows by using the law of iterated expectation and that $E[s_i|\, y_{i0}, D_i, z_i, y_i]\, Z_i' V_1^{-1} e_{1i}$ $= E\left[s_i Z_i' V_1^{-1} e_{1i} \middle|\, y_{i0}, D_i, z_i, y_i\right]$, the third equality sign follows by (81) and (87), and the fourth equality sign holds by the result in (96). This shows that

$$E\left(\frac{\partial l_i(\vartheta)}{\partial \xi_1}\right) = E\left(E\left[\frac{\partial l_i(\vartheta)}{\partial \xi_1} \middle|\, y_{i0}, D_i, z_i\right]\right) = 0 \tag{100}$$

By repeating these arguments we prove the result in (76) for the remaining elements in ϑ. For this purpose we use that

$$\begin{aligned} -T + E\left[\upsilon_{ji}' V_j^{-1} \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] &= 0 && (101) \\ -1 + \frac{1}{\sigma_{j\varepsilon}^2 (1 + T q_j)} E\left[\upsilon_{ji}' J_T \upsilon_{ji} \middle|\, y_{i0}, D_i, z_i\right] &= 0 && (102) \\ p_i - E[s_i|\, y_{i0}, D_i, z_i] &= 0 && (103) \end{aligned}$$

Next, to show (77) we note that for all $t = 1, ..., T$

$$y_{it} \leq \sum_{j=1,2} \left(\left(\rho_j^t + \left(1 + ... + \rho_j^{t-1}\right) \alpha_j\right) y_{i0} + \left(\sum_{s=0}^{t-1} \rho_j^s z_{it-1-s}\right)' \omega_j + \sum_{s=0}^{t-1} \rho_j^s \upsilon_{j,it-s} \right) \tag{104}$$

This implies that $E\left|y_{it}\right|^{k}<\infty$ when $E\left|y_{i0}\right|^{k}<\infty$, $E\left\|z_{it}\right\|^{k}<\infty$ and $E\left|v_{j,it}\right|^{k}<\infty$ for all $t=1,...,T$. According to Assumption 1 and 2 this implies that $E\left|y_{it}\right|^{k}<\infty$ for all $t=1,...,T$ and $k=1,...,6$. In particular, this implies that $E\left\|Z_{i}\right\|^{k}<\infty$ and $E\left\|y_{i}\right\|^{k}<\infty$. Using this we have

$$\begin{aligned} E\left\|\frac{\partial l_{i}\left(\vartheta\right)}{\partial\xi_{1}}\right\|^{2} &= E\left(p_{i}^{*2}\left\|Z_{i}^{\prime}V_{1}^{-1}e_{1i}\right\|^{2}\right)\leq E\left\|Z_{i}^{\prime}V_{1}^{-1}\left(y_{i}-Z_{i}\xi_{1}\right)\right\|^{2} \\ &\leq E\left\|Z_{i}^{\prime}V_{1}^{-1}y_{i}\right\|^{2}+E\left\|Z_{i}^{\prime}V_{1}^{-1}Z_{i}\xi_{1}\right\|^{2}<\infty \end{aligned} \tag{105}$$

The result in (77) for the remaining elements follow by using similar arguments.

To show (78) we show that the conditional mean of the expression in (64) is zero. With respect to the element $p_{i}^{*}H^{\xi_{1}\xi_{1}}$ we use the following. The conditional mean of element (1,1) in the matrix $Z_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$\begin{aligned} & E\left[\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)^{\prime}V_{j}^{-1}\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & E\left[v_{ji}^{\prime}A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}A\left(\rho_{j}\right)v_{ji}\middle|\,y_{i0},D_{i},z_{i}\right]+E\left[Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ & +E\left[2v_{ji}^{\prime}A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & \operatorname{tr}\left\{A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}A\left(\rho_{j}\right)V_{j}\right\}+Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right) \end{aligned}$$

and the conditional mean of element (1,1) in the matrix $Z_{ji}^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$\begin{aligned} & E\left[\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & E\left[v_{ji}^{\prime}A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ & +E\left[2v_{ji}^{\prime}A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & \operatorname{tr}\left\{A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}A\left(\rho_{j}\right)V_{j}\right\}+\operatorname{tr}\left\{A\left(\rho_{j}\right)A\left(\rho_{j}\right)\right\}+\operatorname{tr}\left\{A\left(\rho_{j}\right)\right\}^{2}+Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right) \\ = & \operatorname{tr}\left\{A\left(\rho_{j}\right)^{\prime}V_{j}^{-1}A\left(\rho_{j}\right)V_{j}\right\}+Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}Q\left(y_{i0},z_{i},\xi_{j}\right) \end{aligned}$$

The conditional mean of the matrix obtained by deleting the first row and the first column in the matrix $Z_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$E\left[R\left(y_{i0},z_{i}\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)\middle|\,y_{i0},D_{i},z_{i}\right]=R\left(y_{i0},z_{i}\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)$$

and the conditional mean of the corresponding sub-matrix of $Z_{ji}^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$E\left[R\left(y_{i0},z_{i}\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)\middle|\,y_{i0},D_{i},z_{i}\right]=R\left(y_{i0},z_{i}\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)$$

The conditional mean of the first row and last $k+1$ columns in the matrix $Z_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$E\left[\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)\middle|\,y_{i0},D_{i},z_{i}\right]=Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)$$

and the conditional mean of the corresponding sub-matrix of $Z_{ji}^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}Z_{ji}$ is given by

$$\begin{aligned} & E\left[\left(A\left(\rho_{j}\right)v_{ji}+Q\left(y_{i0},z_{i},\xi_{j}\right)\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & E\left[Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}v_{ji}v_{ji}^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right)\middle|\,y_{i0},D_{i},z_{i}\right] \\ = & Q\left(y_{i0},z_{i},\xi_{j}\right)^{\prime}V_{j}^{-1}R\left(y_{i0},z_{i}\right) \end{aligned}$$

Altogether, this shows that

$$E\left[Z_{ji}'V_j^{-1}Z_{ji}\middle|\, y_{i0},D_i,z_i\right]=E\left[Z_{ji}'V_j^{-1}\upsilon_{ji}\upsilon_{ji}'V_j^{-1}Z_{ji}\middle|\, y_{i0},D_i,z_i\right]\quad\text{for } j=1,2 \tag{106}$$

Using this together with the expression in (65) we obtain

$$\begin{aligned}
&E\left[p_i^*H^{\xi_1\xi_1}\middle|\, y_{i0},D_i,z_i\right]\\
=\;&E\left[s_i\left(-Z_i'V_1^{-1}Z_i+Z_i'V_1^{-1}e_{1i}e_{1i}'V_1^{-1}Z_i\right)\middle|\, y_{i0},D_i,z_i\right]\\
=\;&p_iE\left[-Z_{1i}'V_1^{-1}Z_{1i}+Z_{1i}'V_1^{-1}\upsilon_{1i}\upsilon_{1i}'V_1^{-1}Z_{1i}\middle|\, y_{i0},D_i,z_i\right]\\
=\;&0
\end{aligned}\tag{107}$$

where as above the first equality sign follows by inserting the expression for p_i^* in (80) together with the law of iterated expectation and the second equality sign follows by (81) and (87).

To show the result in (78) for the remaining elements in ϑ we use the following together with the results already obtained

$$E\left[\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=2\operatorname{tr}\left\{I_T\right\}+\left(\operatorname{tr}\left\{I_T\right\}\right)^2=2T+T^2 \tag{108}$$

$$E\left[\upsilon_{ji}'J_T\upsilon_{ji}\upsilon_{ji}'J_T\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=2\operatorname{tr}\left\{J_TV_jJ_TV_j\right\}+\left(\operatorname{tr}\left\{J_TV_j\right\}\right)^2=3\sigma_{j\varepsilon}^4\left(1+Tq_j\right)^2 \tag{109}$$

$$E\left[\upsilon_{ji}'J_T\upsilon_{ji}'\upsilon_{ji}V_j^{-1}\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=\operatorname{tr}\left\{J_TV_j\right\}\left(2+\operatorname{tr}\left\{I_T\right\}\right)=\sigma_{j\varepsilon}^2\left(1+Tq_j\right)\left(2+T\right) \tag{110}$$

$$E\left[\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\upsilon_{ji}'V_j^{-1}A\left(\rho_j\right)\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=2\operatorname{tr}\left\{A\left(\rho_1\right)\right\}+\operatorname{tr}\left\{I_T\right\}\operatorname{tr}\left\{A\left(\rho_1\right)\right\}=0 \tag{111}$$

$$E\left[\upsilon_{ji}'J_TA\left(\rho_j\right)\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=\sigma_{j\varepsilon}^2\left(1+Tq_j\right)\operatorname{tr}\left\{A\left(\rho_j\right)'J_T\right\} \tag{112}$$

$$E\left[\upsilon_{ji}'J_T\upsilon_{ji}\upsilon_{ji}'V_j^{-1}A\left(\rho_j\right)\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]=2\sigma_{j\varepsilon}^2\left(1+Tq_j\right)\operatorname{tr}\left\{A\left(\rho_j\right)'J_T\right\} \tag{113}$$

Again, these results follow by Lemma 1 as $\left(\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right)\sim N\left(0,V_j\right)$. In addition, we have used the result in (91) and that $J_TV_j=\sigma_{j\varepsilon}^2\left(1+Tq_j\right)J_T$ such that $\operatorname{tr}\left\{J_TV_jJ_TV_j\right\}=\sigma_{j\varepsilon}^4\left(1+Tq_j\right)^2\operatorname{tr}\left\{J_T\right\}=\sigma_{j\varepsilon}^4\left(1+Tq_j\right)^2$.

The expression in (66) has mean zero since

$$\begin{aligned}
E\left[\left(-T+\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\right)^2\middle|\, y_{i0},D_i,z_i\right]&=T^2+2T+T^2-2T^2=2T\\
2E\left[T-2\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]&=2T-4T=-2T
\end{aligned}$$

The expression in (67) has mean zero since

$$\begin{aligned}
E\left[\left(-1+\frac{1}{\sigma_{j\varepsilon}^2\left(1+Tq_j\right)}\upsilon_{ji}'J_T\upsilon_{ji}\right)^2\middle|\, y_{i0},D_i,z_i\right]&=1+3-2=2\\
2E\left[1-\frac{2}{\sigma_{j\varepsilon}^2\left(1+Tq_j\right)}\upsilon_{ji}'J_T\upsilon_{ji}\middle|\, y_{i0},D_i,z_i\right]&=2-4=-2
\end{aligned}$$

The expression in (68) has mean zero since

$$\begin{aligned}
\left(T+1\right)E\left[\upsilon_{ji}'V_j^{-1}Z_{ji}\middle|\, y_{i0},D_i,z_i\right]&=0\\
E\left[\upsilon_{ji}'V_j^{-1}\upsilon_{ji}\upsilon_{ji}'V_j^{-1}Z_{ji}\middle|\, y_{i0},D_i,z_i\right]&=0
\end{aligned}$$

The expression in (69) has mean zero since

$$\begin{aligned} -E\left[\upsilon_{ji}' V_j^{-1} A\left(\rho_j\right) \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] &= 0 \\ \frac{1}{\sigma_{j\varepsilon}^2\left(1+Tq_j\right)}\left[\upsilon_{ji}' J_T \upsilon_{ji} \upsilon_{ji}' V_j^{-1} A\left(\rho_j\right) \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] &= 2\,\mathrm{tr}\left\{A\left(\rho_j\right)' J_T\right\} \\ -\frac{2}{\sigma_{j\varepsilon}^2\left(1+Tq_j\right)}\left[\upsilon_{ji}' J_T A\left(\rho_j\right) \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] &= -2\,\mathrm{tr}\left\{A\left(\rho_j\right)' J_T\right\} \end{aligned}$$

The expression in (70) has mean zero since

$$\begin{aligned} T - E\left[\upsilon_{ji}' V_j^{-1} \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] &= 0 \\ E\left[\upsilon_{ji}' J_T \upsilon_{ji} \upsilon_{ji}' V_j^{-1} \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] - (T+2) E\left[\upsilon_{ji}' J_T \upsilon_{ji} \middle| y_{i0}, D_i, z_i\right] &= 0 \end{aligned}$$

With respect to the remaining elements, the result in (78) holds according to the results used to show (76).

The matrix in (78) is positive definite if and only if the elements in the first order derivatives of the log-likelihood function are linear independent. If $\left(\xi_1, \sigma_{1\varepsilon}^2, q_1\right) = \left(\xi_2, \sigma_{2\varepsilon}^2, q_2\right)$ such that $p_i^* = p_i$ for all $i = 1, ..., N$ then according to (54) we have $\frac{\partial l_i}{\partial \gamma} = 0$ for all $i = 1, ..., N$. This means that the standard regularity conditions are not satisfied when the two components in the mixture model are the same. This result is well-known from the literature on mixture models, see Titterington, Smith & Markov (1985) and McLachlan & Peel (2000). In addition, if the variables in z_i are collinear such that the rows in Z_i' are linear dependent for all $i = 1, ..., N$, then the rows in $\frac{\partial l_i}{\partial \xi_j}$ for $j = 1, 2$ are linear dependent for all $i = 1, ..., N$. If the variables in D_i are collinear, then the rows in $\frac{\partial l_i}{\partial \gamma}$ are linear dependent for all $i = 1, ..., N$. Finally, if for any observation (y_i, Z_i) we can choose ξ_1 such that $(y_i - Z_i\xi_1) = 0$ which implies that $\frac{\partial l_i}{\partial \xi_1} = 0$. For any other j where $(y_j - Z_j\xi_1) \neq 0$ we have that as $\sigma_{1\varepsilon}^2 \to 0$ then $p_j^* \to 0$. This implies that $\frac{\partial l_j}{\partial \xi_1} \to 0$ as $\sigma_{1\varepsilon}^2 \to 0$. This means that as $\sigma_{1\varepsilon}^2 \to 0$ we have $\frac{\partial l_i}{\partial \xi_1} \to 0$ for all $i = 1, ..., N$. The possibility of choosing ξ_1 such that $(y_i - Z_i\xi_1) = 0$ for any observation (y_i, Z_i) is ruled out by assuming that the number of rows in Z_i is greater than the rank of Z_i. Altogether, we have shown that the condition in (ii) is satisfied.

(iii) The third order derivatives of the log-likelihood function are obtained by using the expression in (63). As an example, we consider the third order derivative with respect to an element in ξ_1 denoted $\tilde{\xi}$. The corresponding column in Z_i is denoted $\tilde{Z}_i$. We have

$$\begin{aligned} \frac{\partial^3 l_i}{\left(\partial\tilde{\xi}\right)^3} &= p_i^* \frac{\partial^3 \log \phi_i^1}{\left(\partial\tilde{\xi}\right)^3} + \frac{\partial p_i^*}{\partial\tilde{\xi}} \frac{\partial^2 \log \phi_i^1}{\left(\partial\tilde{\xi}\right)^2} + 2p_i^*\left(1-p_i^*\right) \frac{\partial^2 \log \phi_i^1}{\left(\partial\tilde{\xi}\right)^2} \frac{\partial \log \phi_i^1}{\partial\tilde{\xi}} + \left(1-2p_i^*\right) \frac{\partial p_i^*}{\partial\tilde{\xi}} \left(\frac{\partial \log \phi_i^1}{\partial\tilde{\xi}}\right)^2 \\ &= \left(1-2p_i^*\right) p_i^*\left(1-p_i^*\right) \left(\frac{\partial \log \phi_i^1}{\partial\tilde{\xi}}\right)^3 + 3p_i^*\left(1-p_i^*\right) \frac{\partial^2 \log \phi_i^1}{\left(\partial\tilde{\xi}\right)^2} \frac{\partial \log \phi_i^1}{\partial\tilde{\xi}} \\ &= \left(1-2p_i^*\right) p_i^*\left(1-p_i^*\right) \left(\tilde{Z}_i' V_1^{-1}\left(y_i - \tilde{Z}_i\tilde{\xi}\right)\right)^3 - 3p_i^*\left(1-p_i^*\right) \tilde{Z}_i' V_1^{-1} \tilde{Z}_i \tilde{Z}_i' V_1^{-1}\left(y_i - \tilde{Z}_i\tilde{\xi}\right) \end{aligned}$$

such that

$$\left|\frac{\partial^3 l_i}{(\partial\tilde{\xi})^3}\right| \leq \left|\tilde{Z}_i' V_1^{-1}\left(y_i - \tilde{Z}_i\tilde{\xi}\right)\right|^3 + 3\left|\tilde{Z}_i' V_1^{-1} \tilde{Z}_i \tilde{Z}_i' V_1^{-1}\left(y_i - \tilde{Z}_i\tilde{\xi}\right)\right|$$

Since this function is continuous in $\tilde{\xi}$ we can find constants λ_1, λ_2 and λ_3 such that

$$\begin{aligned}\left|\tilde{Z}_i' V_1^{-1}\left(y_i - \tilde{Z}_i\tilde{\xi}\right)\right| &\leq \lambda_1\left|\tilde{Z}_i' y_i\right| + \lambda_2\left|\tilde{Z}_i'\tilde{Z}_i\right| \\ \left|\tilde{Z}_i' V_1^{-1}\tilde{Z}_i\right| &\leq \lambda_3\left|\tilde{Z}_i'\tilde{Z}_i\right|\end{aligned}$$

for $\tilde{\vartheta} \in N(\vartheta)$. This means that for some constants $\tilde{\lambda}_1$, $\tilde{\lambda}_2$, $\tilde{\lambda}_3$ and $\tilde{\lambda}_4$ we have

$$E\sup_{\tilde{\vartheta}\in N(\vartheta)}\left|\frac{\partial^3 l_i}{(\partial\tilde{\xi})^3}\right| \leq \tilde{\lambda}_1 E\left(\left|\tilde{Z}_i' y_i\right|^3\right) + \tilde{\lambda}_2 E\left(\left|\tilde{Z}_i'\tilde{Z}_i\right|^3\right) + \tilde{\lambda}_3\sqrt{E\left(\left|\tilde{Z}_i' y_i\right|^2\right) E\left(\left|\tilde{Z}_i'\tilde{Z}_i\right|^2\right)} + \tilde{\lambda}_4 E\left(\left|\tilde{Z}_i'\tilde{Z}_i\right|^2\right)$$

Using this we have that $E\sup_{\tilde{\vartheta}\in N(\vartheta)}\left|\frac{\partial^3 l_i}{(\partial\tilde{\xi})^3}\right| < \infty$ since $E\left(\left|\tilde{Z}_i'\tilde{Z}_i\right|^3\right) = E\left\|\tilde{Z}_i\right\|^6 < \infty$ and $E\left(\left|\tilde{Z}_i' y_i\right|^3\right) \leq \sqrt{E\left\|\tilde{Z}_i\right\|^6 E\left\|y_i\right\|^6} < \infty$, see also above. The condition in (iii) for the remaining elements in ϑ is shown in a similar manner.

Altogether, we have shown that the density function defined in (12) satisfies Condition 1 and Theorem 1 follows directly. □

B Appendix

Table 4: Log annual earnings

Year	Number of observations	Mean	Standard deviation
1968	262	10.1037	0.3976
1969	288	10.1170	0.4129
1970	308	10.0908	0.4061
1971	339	10.1046	0.4169
1972	369	10.1516	0.4284
1973	401	10.1721	0.4151
1974	427	10.1653	0.4233
1975	455	10.0914	0.4624
1976	490	10.1262	0.4627
1977	523	10.2214	0.4628
1978	562	10.2376	0.4320
1979	562	10.2412	0.4512
1980	562	10.1946	0.4390
1981	562	10.1760	0.4692
1982	562	10.1604	0.4876
1983	525	10.2030	0.4937
1984	488	10.2363	0.4949
1985	463	10.2449	0.5179
1986	449	10.2693	0.5242
1987	428	10.2821	0.5265
1988	407	10.2777	0.5484
1989	393	10.2739	0.5304
1990	374	10.1699	0.5398
1991	358	10.1437	0.5448
1992	333	10.2007	0.5773

Table 5: Parameter estimates from the logistic mixture model (standard errors are in brackets)

Variable	Group 1	Group 2
Lagged log earnings ρ	0.5043 (0.0145)*	0.7060 (0.0151)*
Constant	2.3179 (0.6482)*	0.5728 (0.2611)*
Dummy for high school dropout	-0.1786 (0.0745)*	-0.0168 (0.0317)
Dummy for college graduate	0.0656 (0.0538)	0.0660 (0.0232)*
Linear time trend	-0.0003 (0.0021)	0.0021 (0.0009)*
Age	0.0446 (0.0371)	0.0276 (0.0135)*
Age^2/100	-0.0561 (0.0544)	-0.0432 (0.0197)*
Lagged constant	0.0600 (0.5913)	0.2551 (0.2142)
Lagged dummy for high school dropout	0.0992 (0.0700)	-0.0225 (0.0292)
Lagged dummy for college graduate	0.0901 (0.0507)	0.0192 (0.0210)
Lagged age	-0.0046 (0.0368)	-0.0203 (0.0132)
Lagged age^2/100	0.0066 (0.0553)	0.0348 (0.0198)
Initial log earnings α	0.1854 (0.0240)*	0.2014 (0.0174)*
Short-run variance σ_ε^2	0.1074 (0.0025)*	0.0170 (0.0005)*
$q = \sigma_\alpha^2/\sigma_\varepsilon^2$	0.1886 (0.0263)*	0.2593 (0.0407)*
Mixing weight p_i :		
Constant		0.0544 (0.1841)
Dummy for high school dropout		-0.0049 (0.3215)
Dummy for college graduate		-0.1469 (0.2119)
Log-likelihood: 684.99		

* The parameter is significantly different from zero at the 5% level

Table 6: Parameter estimates from the mixture model where $q_1 = q_2 = 0$ (standard errors are in brackets)

Variable	Group 1	Group 2
Lagged log earnings ρ	0.6723 (0.0110)*	0.8845 (0.0084)*
Constant	1.8021 (0.1702)*	0.5164 (0.0859)*
Dummy for high school dropout	-0.0594 (0.0173)*	-0.0240 (0.0070)*
Dummy for college graduate	0.1077 (0.0125)*	0.0404 (0.0056)*
Age	0.0237 (0.0064)*	-0.0037 (0.0027)
Age^2/100	-0.0297 (0.0081)*	0.0038 (0.0033)
Initial log earnings α	0.1056 (0.0121)*	0.0760 (0.0090)*
Short-run variance σ_ε^2	0.1198 (0.0028)*	0.0191 (0.0006)*
Mixing weight p	0.5161 (0.0234)*	
Log-likelihood: 512.33		

* The parameter is significantly different from zero at the 5% level

References

Alvarez, J., M. Browning and M. Ejrnæs, 2002, Modelling income processes with lots of heterogeneity, CAM Working Paper 2002-01.

Amemiya, T., 1985, Advanced Econometrics (Blackwell).

Andrews, D.W.K. and W. Ploberger, 1994, Optimal test when a nuisance parameter is present only under the alternative, Econometrica 62, 1383-1414.

Abowd, J.M. and D. Card, 1989, On the covariance structure of earnings and hours changes, Econometrica 57, 411-445.

Baltagi, B.H., 1995, Econometric analysis of panel data (Wiley).

Bhargava, A. and J.D. Sargan, 1983, Estimating dynamic random effects models from panel data covering short time periods, Econometrica 51, 1635-1659.

Blundell, R. and S. Bond, 1998, Initial conditions and moment restrictions in dynamic panel data models, Journal of Econometrics 87, 115-143.

Blundell, R.W. and R.J. Smith, 1991, Initial conditions and efficient estimation in dynamic panel data models, Annales d'Economie et de Statistique 20-21, 109-123.

Breitung, J. and W. Meyer, 1994, Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated?, Applied Economics 26, 353-361.

Breitung, J., 1997, Testing for unit roots in panel data using a GMM approach, Statistical Papers 38, 253-269.

Chamberlain, G., 1980, Analysis of covariance with qualitative data, Review of Economic Studies 47, 225-238.

Cramér, H, 1946, Mathematical methods of statistics (Princeton University Press).

Davies, R.B., 1977, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 64, 247-254.

Davies, R.B., 1987, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 74, 33-43.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977, Maximum likelihood estimation from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 1:38.

Hansen, B.E., 1996, Inference when a nuisance parameter is not identified under the null hypothesis, Econometrica 64, 413-430.

Hathaway, R.J., 1985, A constrained formulation of maximum-likelihood estimation for normal mixture distributions, The Annals of Statistics 13, 795-800.

Harris, R.D.F. and E. Tzavalis, 1999, Inference for unit roots in dynamic panels where the time dimension is fixed, Journal of Econometrics 91, 201-226.

Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu, 1999, Bayes estimation of short-run coefficients in dynamic panel data models, in: C. Hsiao et al., eds., Analysis of Panels and Limited Dependent Variables: A volume in Honour of G.S. Maddala (Cambridge University Press) 268-296.

Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu, 2002, Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods, Journal of Econometrics 109, 107-150.

Im, K.S., M.H. Pesaran and Y. Shin, 2003, Testing for unit roots in heterogeneous panels, Journal of Econometrics 115, 53-74.

Kiefer, N.M., 1978, Discrete parameter variation: Efficient estimation of a switching regression model, Econometrica 46, 427-434.

Kiefer, J. and J. Wolfowitz, 1956, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Annals of Mathematical Statistics 27, 887-906.

McCurdy, T.E., 1982, The use of time series processes to model the error structure of earnings in a longitudinal data analysis, Journal of Econometrics 18, 83-114.

McLachlan, G. and D. Peel, 2000, Finite mixture models (Wiley).

Neyman, J. and E. Scott, 1948, Consistent estimates based on partially consistent observations, Econometrica 16, 1-32.

Nielsen, B. and A. Rahbek, 2000, Similarity issues in cointegration analysis, Oxford Bulletin of Economics and Statistics 62, 5-22.

Pesaran, M.H., Y. Shin and R.P. Smith, 1999, Pooled mean group estimation of dynamic heterogeneous panels, Journal of the American Statistical Association 94, 621-634.

Pesaran, M.H. and R. Smith, 1995, Estimating long-run relationships from dynamic heterogeneous panels, Journal of Econometrics 68, 79-113.

Policello, G.E., 1981, Conditional maximum likelihood estimation in Gaussian mixtures, in C. Taillie et al., eds., Statistical Distributions in Scientific Work, Vol.5, 111-125.

Quandt, R.E., 1972, A new approach to estimating switching regressions, Journal of the American Statistical Association 67, 306-310.

Rahbek, A. and N. Shephard, 2002, Inference and ergodicity in the autoregressive conditional root model, working paper 2002-W7, Nuffield College, Oxford University.

Redner, R.A. and H.W. Walker, 1984, Mixture densities, maximum likelihood and the EM algorithm, SIAM Review 26, 195-239.

Robertson, D. and J. Symons, 1992, Some strange properties of panel data estimators, Journal of Applied Econometrics 7, 175-189.

Titterington, D.M., A.F.M. Smith and U.E. Markov, 1985, Statistical analysis of finite mixture distributions (Wiley).

Wong, C.S. and W.K. Li, 2000, On a mixture autoregressive model, Journal of the Royal Statistical Society B 62, Part 1, 95-115.

Wong, C.S. and W.K. Li, 2001, On a logistic mixture autoregressive model, Biometrika 88, 833-846.

Wooldridge, J.M., 2002, Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity, CEMMAP working paper CWP18/02.

Wooldridge, J.M., 2002, Econometric analysis of cross section and panel data (MIT Press).